

THE CONSTRUCTION AND USE OF ACHIEVEMENT EXAMINATIONS

A Manual for Secondary School Teachers

PREPARED UNDER THE AUSPICES OF A COMMITTEE OF THE
AMERICAN COUNCIL ON EDUCATION

*Herbert E. Hawkes, Chairman
Algernon Coleman, John Lester, E. F. Lindquist, John
A. Long, R. W. Tyler, Ben D. Wood, George F. Zook*

EDITORS

HERBERT E. HAWKES, E. F. LINDQUIST, C. R. MANN



GEORGE G. HARRAP & CO. LTD.
LONDON BOMBAY SYDNEY

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

PROBABLY more than a million people in the United States are officially charged with responsibility for writing examination questions. College professors are but a small proportion — about 5 per cent. Besides these, there are school teachers of all grades, civil service examiners, state examiners for licensing doctors, pharmacists, and other professional men, bar examiners, and many others.

Examinations are not only used extensively; they vitally affect the lives and fortunes of millions of our people. Careers are sometimes determined by them. They are powerful instruments for promoting or for retarding realization of that great American aspiration to give everyone his utmost chance. Therefore, it is not enough merely to make sure that they are valid and fairly administered. We must make them reveal the salient factors involved as accurately and as intelligently as possible.

Fifty years ago examining was a highly individualistic affair. Each examiner framed his own questions, read the papers himself, and made his personal appraisal of the results. This simple procedure was adequate while our civilization was developing through the pioneer stage. But, as wealth increased, more young people pressed on in increasing numbers to colleges and professional schools, and the need for reliable and discriminating examinations became acute. Hence, the gradual evolution of examining boards, of committees of readers, of improved school records and accrediting systems, and, more recently, of new-type standardized objective tests. These, and many other trends, are both demanding and tending to produce greater reliability and validity in examinations and tests.

PREFACE

At the same time, the purposes and functions of examinations and tests have been clarified and extended. Formerly, the main if not the only purpose of examinations was conceived to be that of separating the sheep from the goats, as in the case of qualifying examinations for admission to one of the learned professions, or qualifying examinations for a college or university degree, or for admission to college or university. These are still very important functions of examinations, but recently increasing emphasis has been placed upon examinations as means for improving instruction, and as instruments for securing information that is indispensable for the constructive educational guidance of pupils. The concluding chapter of this volume is largely an exposition of the thesis that the emergent major rôle of comparable examinations in American education is that of helping to ascertain and meet the needs of individual pupils in our very heterogeneous school and college populations.

Before examinations can be expected to serve their purpose completely, it is of the utmost importance that examiners and test makers clearly understand what they are trying to do and what the results mean. Of course, no one can devise a proper examination or test without having some idea of what he wants his test to appraise. But with the average examiner these guiding ideas are usually so vague and abstract that the resulting questions represent just one individual's guess. Hence, the marked trend in recent years toward clarification and specification of the objective results which schooling aims to help people achieve and which examinations are designed to measure.

Particularly during the past fifteen or twenty years, many organizations as well as individuals have been active in the attempt to state the objective results desired by educational effort, and in the effort to devise accurate means of evaluating

PREFACE

such results. Colleges and universities, state boards of education, teachers' associations, various examining bodies have all been at work. The result is a vast store of information, most of which never reaches those who are actually doing the classroom teaching and who therefore ought to be at the center of gravity of the whole movement. This volume is an attempt to bring to the attention of teachers and administrators, so far as can be done in so narrow a compass, a survey of the principles that lie at the basis of any system of examination making that deserves the name, and to describe in some detail sound methods of test making in the various subject-matter fields.

About five years ago, the proposal was made that such a volume could best be prepared with the collaboration of a number of experts in the theory and practice of examining rather than by any one individual. In order to provide for the conferences and collaboration that seemed necessary for such an enterprise, an appropriation was sought and generously granted by the Carnegie Foundation for underwriting the preparation of the book.

The volume is aimed particularly at the needs of the classroom teacher in schools and in the first two years of college. It is hoped that it will serve not only to assist the teacher in the actual process of producing examinations that will accomplish their purpose better than would otherwise be the case, but also to ground him in the type of objective results and principles that lie at the basis of examination making.

The latter of these purposes is fully as important as the former. The task of achievement test construction, as this volume should make fully apparent, has become increasingly technical and specialized in character. Few teachers can take the time, even though they have acquired the necessary technical skill, to prepare for themselves more than a small part of

PREFACE

the examination materials required in instruction. More reliance must be placed, in the production of test materials, upon specially organized test building agencies, such as the Cooperative Test Service and the various state and regional testing programs. The availability of expertly constructed materials, however, does not relieve the classroom teacher of the responsibility of becoming intimately familiar with the procedures employed in good test construction, or of knowing in detail the desirable characteristics of a good examination. The classroom teacher must learn how to build good tests, not so much in order that he may construct his own examinations, as that he may make the most effective use of and best interpret the results obtained from examinations constructed by others.

It may appear that a considerable portion of the volume is devoted to the so-called new-type test, while the average teacher is more concerned with the setting of the more conventional style of examinations. Notwithstanding this apparent emphasis, it will be observed that the principles which lie at the basis of the entire treatment are applicable to all examinations, whether they be of the more common subjective type, or the comprehensive examination, or the new-type objective kind of test. Furthermore, the volume should serve the purpose of deterring many an enthusiast from attempting the construction of examinations which he has neither the time nor the technical training to prepare properly. The opinion is stated again and again that the task of preparing a set of new-type examinations that are valid and comparable is a matter demanding the rarest technical skill, and that if such examinations are desired, they might well be secured from those who are experts in their construction. It is plainly common sense so to use tests that they help rather than hinder the evolution of the child.

The omission of a chapter explicitly devoted to the so-called

PREFACE

comprehensive examination may possibly cause surprise. The publication of the two volumes on this phase of examination making by Dr. E. S. Jones of the University of Buffalo seemed to render the inclusion of such a chapter in this book unnecessary. Teachers who are interested in this aspect of the subject are referred to Dr. Jones's work.

Special acknowledgments are due to Professor J. J. Coss for his help in getting this project organized and to Dr. E. R. Smith and Dr. Ben D. Wood for their extensive help to the authors throughout the preparation of the book.

CONTENTS

Part I

GENERAL CONSIDERATIONS

I. IDENTIFICATION AND DEFINITION OF THE OBJECTIVES TO BE MEASURED	3
R. W. TYLER, <i>Ohio State University</i>	
II. THE THEORY OF TEST CONSTRUCTION	17
E. F. LINDQUIST, <i>State University of Iowa</i>	
III. THE CONSTRUCTION OF TESTS	107
E. F. LINDQUIST	

Part II

EXAMINATIONS IN MAJOR SUBJECT FIELDS

IV. EXAMINATIONS IN THE SOCIAL STUDIES	163
HOWARD R. ANDERSON, <i>State University of Iowa</i>	
V. EXAMINATIONS IN THE NATURAL SCIENCES	214
FRED P. FRUTCHEY and R. W. TYLER, <i>Ohio State University</i>	
VI. EXAMINATIONS IN THE FOREIGN LANGUAGES	264
ALGERNON COLEMAN, <i>University of Chicago</i>	
VII. EXAMINATIONS IN MATHEMATICS	337
JOHN A. LONG, <i>University of Toronto</i> , HAROLD T. LUNDHOLM, <i>Blake School, Minneapolis</i> , and EUGENE R. SMITH, <i>Beaver Country Day School, Massachusetts</i>	
VIII. EXAMINATIONS IN ENGLISH	381
JOHN A. LESTER, <i>formerly Hill School, Pennsylvania</i> , and E. F. LINDQUIST	

CONTENTS

Part III

THE FUNCTIONS AND LIMITATIONS OF EXAMINATIONS

IX. THE USES AND ABUSES OF EXAMINATIONS .	443
MAX MCCONN, <i>Lehigh University</i>	
BIBLIOGRAPHY	479
LATIN ACHIEVEMENT TESTS	490
INDEX	493

PART I
GENERAL CONSIDERATIONS

CHAPTER I

IDENTIFICATION AND DEFINITION OF THE OBJECTIVES TO BE MEASURED

DISCUSSIONS of methods to be followed in constructing examinations are usually arguments for or against new-type tests as opposed to essay examinations. Apparently the question of the form of the examination has engrossed the attention of teachers to the exclusion of other questions. The situation is very similar to an argument over the merits of particular automobiles in which all of the discussion centers around the advantages of a coupé in contrast to a roadster. The type of body is, of course, a factor in selecting an automobile, but the first question ought to be concerned with the effectiveness of the motor. An automobile is purchased for certain uses, and the value of the machine must be judged in terms of those uses. In the same way, a test or examination is made for certain uses, and the problem of first concern to teachers should be the value of particular examinations for the uses to which they are to be put.

What are the uses which an examination must serve? Tests are given in order to discover the difficulties which the class is having so that the teaching emphasis may be directed toward overcoming these difficulties; to discover the difficulties of individual students so that they may direct their study more wisely; to give the instructor an estimate of the effectiveness of his own instruction. Increasingly, tests are being used to determine the effectiveness of particular teaching procedures, or particular methods of selecting and organizing course materials. All of these uses require tests which really show how well the students are progressing in their school work.

Hence, a satisfactory test or examination must first be one which actually gives us evidence of the amount of progress which students are making.

What constitutes progress in school work? It is certainly true that every change which takes place in a student during the time he is in school cannot be considered progress. During the time he is in school the student may grow taller, he may grow fatter, he may acquire a new slang vocabulary, his voice may change, but we do not consider these as evidences of his progress in his school work. Each subject which is taught is offered with the expectation that students who take this subject will undergo certain desired changes as a result of the course. In algebra, for example, it is expected that students will acquire a certain understanding of the meaning of abstract number, and that they will become somewhat skillful in solving numerical problems. These changes which we expect to take place in the student are the objectives of the subject. It is apparent that a satisfactory test in algebra is one which shows us the degree to which students are reaching these objectives, i.e., the degree to which they have acquired an understanding of the meaning of abstract number, and the degree to which they have become skillful in solving numerical problems. In similar fashion, every subject offered involves certain objectives which it is hoped students will reach as a result of instruction in this subject. A satisfactory test or examination in any subject is an instrument which gives evidence of the degree to which students are reaching the objectives of teaching.

This point of view considered, it becomes necessary to enlarge the common conception of a test or examination. Many people have limited the concept of an examination to a paper and pencil test. This is obviously a harmful limitation. Sometimes the best way to get evidence of the desirable changes

OBJECTIVES TO BE MEASURED

which are taking place in students is through observation, or by other means. To use only paper and pencil tests might seriously restrict the opportunity for determining the progress students are making. An examination then is essentially a means for getting valid evidence of the degree to which students have attained the desired objectives of instruction. In most subjects there are several important objectives which students and instructors are trying to reach. Hence, a satisfactory examination must correspondingly provide evidence of the degree to which students are reaching each of these important objectives. One major defect of typical examinations has been the fact that they have given evidence with reference to only a limited number of objectives and have not adequately indicated the degree to which students were attaining all of the desired outcomes of instruction.

The importance of covering all of the significant objectives in a total examination program can best be shown by illustration. In chemistry the objectives which instructors are commonly trying to reach include teaching students to acquire a fund of important facts and principles; to understand the technical terms commonly appearing in chemical publications; to be able to apply important chemical principles to appropriate situations; to express chemical reactions by means of equations involving chemical symbols and formulae; to be skillful in certain laboratory techniques. Any adequate examination program for chemistry will provide means for discovering how far each of these objectives is being attained. Tests need to be included which will indicate how well students are acquiring these important facts and principles; how well they understand the technical terms commonly appearing in chemical publications; how well able they are to apply important chemical principles to appropriate situations; how satisfactorily they can express chemical reactions by the use

GENERAL CONSIDERATIONS

of equations; how skillful they are in the essential laboratory techniques. Obviously, evidence of all of these attainments cannot be had from a single examination, but an inclusive examination program should cover all of the important objectives. Some of these attainments can be determined by means of paper and pencil tests with which everyone is more or less familiar. Others would need to be tested by different devices. In order to discover how skillful the students have become in the essential laboratory techniques, it is probably necessary to set the students at work on certain laboratory problems and to evaluate their skill by means of observation and by checking the outcome of the laboratory exercises.

The variety of examinations necessitated by a variety of objectives may also be illustrated in the subject of English. It is probable that among important objectives in this field would be included teaching students to use correct English; to write effectively; to be familiar with significant literature; to be able to evaluate various types of literary productions; to appreciate good literature. A satisfactory program of examinations in English will therefore include tests which reveal the students' ability to use correct English, their ability to write effectively, their familiarity with significant literature, their skill in evaluating literary productions, and their appreciation of good literature. This means again that the examination program involves a variety of testing techniques. Skill in written composition can be judged by the use of paper and pencil tests. The use of correct oral English may need to be evaluated by a different device. To discover how well they have learned to appreciate good literature would probably require still other devices, as, for example, one which would indicate the literary preferences of the students.

The importance of a total examination program which provides evidence of the students' attainments in all the aspects of

the course is not always recognized. When we compare the tests and examinations in common use with the objectives of the courses, it is evident at once that these tests do not show us how well the students are attaining all of these objectives. In many subjects, the typical tests and examinations give us evidence only of the progress students are making in acquiring facts and in understanding the meaning of technical terms in the field. Rarely do we find students tested on such objectives as their ability to utilize scientific method, the consistency of their points of view, their skill in laboratory work. To this criticism of the inadequacy of typical tests, the answer is sometimes made that the acquisition of information is basic to all other objectives. It is claimed that one cannot think without facts and that the test which reveals the degree to which students have acquired important facts indirectly constitutes a test of all of the objectives of instruction. This claim, however, is not justified. In the botany and zoology classes at the Ohio State University, comparisons have been made of the records of the students' grades on tests which show the degree to which they have acquired important facts and on tests which indicate how well they are able to apply principles to new situations. The results are by no means identical.² It is found that many students who have acquired a large number of facts are unable to apply these facts to new situations. Similarly, in history, many students have acquired important facts and yet are unable to apply these facts in interpreting contemporary events. In English many students have been found who can write effective compositions but who cannot use oral English effectively. We do not have a complete picture of the progress students are making when we depend only upon tests of a limited number of objectives.

These subjects are but illustrations of the situation pre-

² None of the correlations are above .40 and most of them are about .25.

GENERAL CONSIDERATIONS

vailing in every field. Because of the importance of having an examination program which gives evidence of the degree to which students are reaching each of the significant objectives of the subject, an essential step in planning an examination program is to formulate in a clear and understandable fashion the important objectives which the instructor is trying to reach. This formulation then becomes the comprehensive plan against which the various tests are checked to be sure that the total examination program includes devices for determining the degree to which students are attaining each of these objectives.

Two problems are usually involved in formulating the objectives of a particular course. One is to get a list of objectives which is reasonably complete, that is, which includes all of the important objectives to be reached. The other is to state the objectives in such clear and definite terms that they can serve as guides in the making of the examination questions. Many statements of objectives are so vague and nebulous that, although they may sound well, they prove to be glittering generalities which are of little use in making examinations.

In making a list of objectives for a course, one procedure commonly followed is to begin with the general function or purpose of the subject and to analyze this into its several aspects or sub-functions. Another method is to begin with the content of the course and with reference to each topic ask the questions: What is the purpose of this topic? What do I expect students to get from this topic? In most cases it is necessary to use a combination of both procedures in order to get a relatively complete list of important objectives and in order to clarify the meaning of each objective.²

² The validation of the objectives and of the course content is an important step in curriculum construction but is not treated here, since it is assumed that this curriculum problem has been attacked before test construction is begun.

OBJECTIVES TO BE MEASURED

This combination of methods can be illustrated by the procedures followed by the department of zoology in a certain university in formulating the objectives for the elementary course. This department recognized two major functions or purposes of the elementary course. One of these was to teach the student a fund of important zoological information; the other was to teach the student to use scientific method in zoology. Beginning with these accepted major functions, the instructors in the department first analyzed them into several sub-functions, i.e., they broke up the general objectives into the several more definite objectives which these general objectives included. Upon analysis, the general informational objective was split up into the following: the recall of important specific facts; the memory of general zoological principles; and the recognition of the meaning of common technical terms found in zoological publications. The use of scientific method, the instructors decided, meant the ability to formulate reasonable generalizations from experimental data, the ability to plan satisfactory experiments to test promising hypotheses in zoology, and the ability to apply significant zoological principles to situations new to the students. By means of an analysis of the two general objectives, the department was thus able to formulate these six more definite objectives.

To check the completeness of these six objectives the instructors then took up, topic by topic, the content of the elementary course, asking themselves in connection with each topic what they expected students to get from the topic. Three objectives were added as a result of this analysis: skill in the laboratory techniques of dissection and the use of the microscope, the ability to report the results of experiments in effective English, and familiarity with sources of information on zoological problems.

In order to make a list of major objectives usable in building

GENERAL CONSIDERATIONS

examinations, each objective must be defined in terms which clarify the kind of behavior that the course should help to develop among the students. That is to say, a statement is needed which explains the meaning of the objective by describing the behavior we can expect of persons who have attained it. ("Behavior" is here used to mean any sort of appropriate reactions of students — mental, physical, emotional, and the like.) For example, in the case of the zoology course, the first objective was to teach students to recall important specific facts. This objective was analyzed by defining expected pupil-behavior in the following terms: To remember and state these facts without having to look them up at the time, and to recognize misconceptions which are commonly mistaken for zoological facts. The analysis of this objective also required a definite statement of the important zoological facts which students are expected to remember and the misconceptions which are commonly mistaken for zoological facts.

Similarly, in analyzing the objective of skill in certain significant laboratory techniques, it was necessary to list the types of dissections which students should be able to make, the kinds of microscopic mounts which they should be able to prepare, and typical objects which they should be able to find under a microscope.

When all these objectives had been defined in terms of behavior and the necessary lists of facts, principles, terms, experiments, and the like had been made, the instructors had all of the basic material which they needed in constructing examinations to cover the major objectives of the zoology course. This procedure suggests the usefulness of the definitions of objectives in terms of behavior. The customary method of analyzing a course as a preliminary step to making examinations has been to analyze only the content of the course. The definition of objectives in terms of expected behavior

OBJECTIVES TO BE MEASURED

differs from the analysis-of-content method, as may easily be seen. Textbooks in zoology, for example, describe certain experiments that have been performed to show that a frog responds to light stimuli in his feeding reactions. A textbook analysis would include the description of these experiments, but on the usual basis of test construction it would be assumed that the student is expected to remember these descriptions. An examination would then be constructed which would disclose whether or not the student remembers the details of these experiments. In contrast, a definition of objectives in terms of student behavior does more than indicate the content to be covered. It defines the reactions which a student is expected to make to this content. Thus, the fourth objective for zoology was defined as the ability to formulate in his own words as complete a generalization as is justified by the data presented, when the student is given the results of a zoological experiment. It is expected, not that a student will remember the details of an experiment, but that upon being given the details he will be able to interpret the experiment for himself. Obviously, with the same content it makes a great deal of difference whether the examination is to test the student's memory of the facts or his ability to interpret them when they have been presented to him.

The lists of facts, principles, terms, experiments, and the like are also necessary parts of the basic materials. Some time is required to collect these basic lists, but when they have been assembled they serve as a reservoir of materials for making new examinations and ultimately save a great deal of time. Furthermore, many of these lists are very helpful means of checking the content of the course and are useful as guides in preparing lectures, in planning laboratory work, and in outlining other class assignments.

GENERAL CONSIDERATIONS

Many subjects, especially the sciences, are constantly changing with the discovery of new facts and principles. Hence, it is necessary to prevent the reservoir of basic materials from becoming static and thus crystallizing what ought to be a developing course. It is, therefore, desirable to provide for an annual review of each of the basic lists by instructors in the department. As each list is examined, items are deleted which have been rendered obsolete by the new developments in the field and other items are added to cover these new developments. The revision requires relatively little time and helps to insure a continuously appropriate set of materials from which examinations may be quickly made.

This illustration taken from the work in zoology is suggestive of the value of formulating and analyzing the objectives of a course in order to make examinations which are satisfactory and relatively complete. Similar procedures can be appropriately used in other fields of subject matter. In a foreign language course, for example, the objectives might include the ability to comprehend the meaning of selections written in the foreign language; the ability to understand oral expression in the foreign language; the ability to pronounce orally words, sentences, and paragraphs in the foreign language; the ability to compose effective written expression in the foreign language; the ability to compose effective oral expression in the foreign language; knowledge of the grammar of the foreign language; an understanding of the important vocabulary in the foreign language; a knowledge of the art, literature, and customs of the people whose language is being studied.

Further analysis is necessary in order to make this list of objectives usable in building examinations. For example, the first objective might be analyzed as follows: the nature of the reading ability should be defined, and a collection should be made of selections written in the foreign language which are

OBJECTIVES TO BE MEASURED

new to the students and which they should be able to comprehend. This collection would serve to define the kinds of reading material to be covered in the examinations in terms of narratives, expositions, descriptions, and the like. In making the collection, care should be exercised to have an appropriate vocabulary included and to obtain selections involving ideas of the appropriate difficulty for the class. By analyzing each of these objectives of the foreign language course in a similar manner the basic materials are obtained from which examinations can easily be made. It is readily apparent that the procedures of formulating and analyzing the major objectives are desirable for any course and are invaluable when making a comprehensive program of examinations.

This statement of objectives represents the general purposes of the course. In appraising the progress of each student adequate consideration should be given to his individual pattern of desirable educational goals.³ The objectives of the course do not represent points to be reached by all students but rather directions in which students may progress. This is an important distinction.

When objectives are conceived as uniform goals to be attained by all students, teaching tends to become an attempt to maintain a lock-step march to these goals, while testing is used to discover whether the students have reached the goals. Such a conception omits the vast array of facts regarding individual differences. Individuals differ not only in rate and methods of learning but in interests, needs, and potential abilities. How far each student may be expected to progress toward any objective varies with his needs, his interests, and those abilities of his which are involved in this progress. The

³ For a description of a plan of education emphasizing individual goals see Paul F. Voelker and others, "A Program of Demonstration and Research," *Educational Record*, vol. XVI, no. 2 (April, 1935), pp. 207-216.

GENERAL CONSIDERATIONS

several objectives of a course thus become directions in which the course may help students to progress. Some of these outcomes are of such significance to all members of our society as to be well-nigh universal educational objectives. Some of these outcomes, however, are highly individual. Correspondingly, some students will go farther in some directions, while others will make greater progress in other directions. The relative emphasis given to the various objectives will differ from one student to another. In this sense, objectives become individualized as do teaching and learning procedures.

It is much easier to accept the median achievement of a group of students as the goal for each person than it is to try to formulate a suitable individual pattern of goals. Hence, the evaluation of each student's progress in terms of objectives appropriate for him is rarely made. This difficulty can be overcome by helping the student periodically to state his goals and to revise his previous statements in the light of continually accumulating evidence of his needs, abilities, and achievements. In this way tentative individual objectives are validated.

To appraise each student's progress with reference to his goals is not as simple as is the common appraisal procedure of adding together a series of numerical scores. The proper conception of evaluation eliminates purely mechanical appraisal and substitutes judgment and thoughtful consideration. However, this does not imply intuitive appraisal but demands valid judgments based upon the careful collection of comprehensive evidence regarding student progress.

This discussion demonstrates the fact that the first steps in examination building are those involved in determining objectives. It is usually necessary to consider first the major functions or purposes of the course, breaking those up into several more definite major objectives. To make this list of

OBJECTIVES TO BE MEASURED

major objectives more nearly complete the content of the course is then examined, topic by topic, to discover why the topic has been included in the course, that is, to state the things which students are expected to gain from studying the topic. This examination of individual topics usually suggests additions to the list of major objectives. The purpose of these steps is to obtain a relatively complete list of the most important objectives. After the list has been formulated, an analysis is then made of each objective to give it definite meaning by defining it in terms of the behavior expected of students, and to obtain a comprehensive statement of the specific elements involved in the objective.) This analysis provides the basic material from which a complete program of examinations may be constructed. Furthermore, a plan of periodic review and revision of these basic materials prevents crystallization of the course and of the examinations. Finally, each objective needs to be conceived as a direction in which the course may help one or more of the students to progress. The progress of each student may then be evaluated in terms of goals appropriate for him.

QUESTIONS FOR DISCUSSION

1. Show by illustration the difference between objectives which are expressed in terms of content to be covered and those expressed in terms of changes in behavior.
2. How specific should be the formulation of objectives for tests? Give some illustrations of objectives which are too general for test purposes, some which are too specific, and some which are of the desired degree of specificity. What factors need to be considered in determining the degree of specificity desired in a formulation of objectives for tests?
3. To what degree should formulations of educational objectives cut across subject-matter lines? That is, under what conditions is it desirable to formulate an educational objective which in-

GENERAL CONSIDERATIONS

- volves two or more subjects? Give some illustrations of some desirable educational objective which involves more than one subject.
4. Criticize each of the following statements of objectives, indicating in what respects it is satisfactory and in what respects it needs improvement.
 - (a) Ability with graphs.
 - (b) Discriminating tastes in reading.
 - (c) Skill in laboratory techniques.
 - (d) Familiarity with the common vocabulary of French.
 - (e) Ability to interpret social-science data intelligently.
 - (f) Habit of selecting meals wisely.
 - (g) Ability to apply important biological principles in explaining common biological phenomena.
 - (h) Understanding of the means by which the economic functions of our society are performed.
 5. Select a field in which you are interested and formulate a statement of educational objectives which might guide the construction of tests for this field.

CHAPTER II

THE THEORY OF TEST CONSTRUCTION

INTRODUCTION

THE constructor of any test of educational achievement is confronted with two major problems. The first of these is the problem of *what* to measure. This problem involves the determination of the general and specific, or ultimate and immediate, objectives of instruction, described in terms of the specific changes (in skills, abilities, attitudes, information, understanding, appreciations, etc.) which it is hoped have been produced in the learner, and the construction or selection of the specific learning materials and situations related to those objectives, on the basis of which test items may be constructed. This problem is essentially one of curriculum construction, for which the subject-matter expert and the curriculum builder rather than the test technician or classroom teacher are primarily responsible, but it is fundamental to adequate test construction and has therefore been given extensive consideration in the preceding chapter. Obviously, until the field of achievement in which measurement is to be attempted has been specifically and authoritatively described, adequate test construction in that field is impossible. It may be noted, in this connection, that perhaps the most serious of the weaknesses which characterize achievement tests now being constructed are basically weaknesses of the curriculum rather than of the test techniques themselves.

The second major problem is that of *how* to measure. After having determined what is to be measured, i.e., after having described the field of achievement in which measurement is to be attempted, the test constructor must decide:

GENERAL CONSIDERATIONS

1. How to select a sampling of the elements of that field (skills, abilities, information, understanding, appreciations, etc.) to constitute the basis of the individual test items;
2. Which testing techniques, or types of test exercise, are best adapted to these elements, individually and collectively;
3. How to phrase, arrange, or present each individual item;
4. How to assemble the items into a complete test or series of tests;
5. How to administer the test;
6. How to evaluate performance on the test, i.e., how to score it and how to interpret the scores;
7. How to evaluate the test itself, i.e., how to determine its validity.

These divisions will each present many minor problems which will vary considerably in nature and importance from one field of subject matter to another. There are, however, a number of problems, principles, and techniques which are fundamental in test construction in all fields and which may advantageously be taken up independently in advance of any consideration of problems unique to the separate fields. The purpose of this chapter, therefore, is to deal with certain aspects of the whole problem of *how to measure* which are common to the major fields of subject matter in secondary schools and colleges, and to present general rules or suggestions for test construction. Supplementary discussions and specific suggestions for each of the various subject-matter fields will be found in the later chapters.

This chapter will be further restricted in scope in that it will deal only with the written examination. This restriction does not imply, of course, that the written examination is

adequate for the measurement of all types of educational achievement. There are many other techniques — including the recitation, the oral examination, the rating scale, the interview, the performance test, the interest questionnaire, the attitude scale, and general observation of student behavior — which have a significant place in the whole process of appraisal and which are seriously in need of further development. Limitations in space, however, as well as the unsatisfactory state of present knowledge concerning these techniques, demand that they be omitted from the present discussion.

In its treatment of the written examination, furthermore, this chapter will be devoted primarily, at least so far as space is concerned, to tests of the type which may be scored objectively. The reasons for this allocation of space are: (1) Considerably more energy and talent have recently been devoted to the study and improvement of the objective than of the essay type of examination. Consequently, the present amount of detailed and dependable knowledge concerning the possibilities, limitations, and methods of construction is now much larger for the objective than for the essay test, while the recency and relatively technical nature of available information concerning objective test construction make it most in need of dissemination and elucidation. (2) The recent widespread and rapid growth of state-wide or regional cooperative testing programs, in which the use of the objective type of test is essential to comparability of results, has increased the need for a better understanding of this type of test on the part of all concerned. And (3) many of the principles basic to the construction of written examinations of all types can be most clearly and conveniently illustrated in the case of the objective test.

It is definitely not implied, however, that the value or place of the essay examination in the whole program of

GENERAL CONSIDERATIONS

measurement is any less significant than that of the objective test. There is, indeed, no need for recognizing any general conflict between these two types. No good purpose can possibly be served by arguing their *general* advantages and disadvantages, while much harm can thereby be done by appealing to established prejudices. The intelligent point of view is that which recognizes that whatever advantages either type may have are *specific* advantages in *specific* situations; that while certain purposes may be best served by one type, other purposes are best served by the other; and, above all, that the adequacy of either type in any specific situation is much more dependent upon the ingenuity and intelligence with which the test is *used* than upon any *inherent* characteristic or limitation of the *type* employed.

THE IMMEDIATE OBJECTIVES OF THE SCHOOL EXAMINATION

Written examinations and the measures secured from them are used in the school to serve a wide variety of purposes, but for the great majority of such examinations the basic or immediate objectives are only two in number. One of these objectives of testing is to *rank* the students tested in the order of their *total* achievement in a given field of subject matter, or within a specified division or portion of that field. The other immediate objective is to discover specific weaknesses, errors, or gaps in the student's achievement. Nearly all of the ultimate uses of testing and of test results — in the improvement of instruction and the motivation of learning, in promotion and grading, in remedial instruction, in supervision and administration, in sectioning, in educational guidance, etc. — depend directly upon the effectiveness of the tests used in relation to one or both of these immediate objectives.

The importance of this discrimination in the immediate objectives of testing becomes apparent upon consideration of the nature of test validity. The most important — the all-important — characteristic of any test is its validity. The validity of the test depends upon the effectiveness with which it measures that which it is *intended* to measure, or, otherwise stated, upon the effectiveness with which it accomplishes the purpose it is intended to accomplish. The mistake is frequently made of describing a test as “valid” or “invalid” in general, implying an “all-or-none” characteristic with no specific reference. Validity, on the contrary, is a highly specific concept, and refers to something in which tests differ only in degree. If a test is “valid,” it is valid for *a given purpose*, with a given group of pupils, and it is valid only to the degree that it accomplishes that specific purpose for that specific group. It is meaningless to speak of any given test as being valid or invalid apart from any consideration of the purpose it is intended to serve or of the group to which it is to be given. A test may serve a given purpose effectively with one group of pupils, but fail seriously to accomplish the same purpose with another group. For a given group, the same test may be highly valid for one purpose, almost completely lacking in validity for another, and may possess intermediate degrees of validity for still others. A given test, for example, may possess high validity for the purpose of ranking a group of high-school pupils in order of their total achievement in a general course in United States history, but may be too easy to discriminate between college students in the same subject. It may possess a lower but still satisfactory validity for measuring general achievement in a high-school course in “The economic history of the United States,” may have some validity but to a less degree for the purpose of predicting probable future success in a high-school course in English

GENERAL CONSIDERATIONS

history, may be even less valid for the purpose of disclosing specific deficiencies in achievement in United States history, still less valid for providing an index of the general intelligence of the pupils tested, and practically invalid as a basis for assigning grades in manual training. Statements concerning the validity of an educational test should therefore always be accompanied by a statement of the purpose to which that validity refers and by a description of the group of pupils for which the test is intended. If a test is to be made most valid for one purpose, its validity for other purposes must often be sacrificed. An "all-purpose" test can perform none of its many purposes effectively.

These considerations are of special importance with reference to the two immediate objectives of the school examination that were mentioned earlier. Unfortunately there are few instances in which the same test may accomplish simultaneously both of these immediate objectives with anything approaching maximum effectiveness. Tests that are highly effective for ranking pupils in order of total or general achievement are often of little value as diagnostic instruments, while a test that is to be significantly diagnostic within a given field must usually be made too long and unwieldy to be practicable as a general achievement test. For example, the mixed-error proof-reading type of test in English (as used in the Cooperative English Test, Series 2) is highly valid for measuring total or general achievement in the mechanics of writing, but it has little or no validity for the purpose of discovering specific deficiencies or habitual errors in the writing of an individual pupil. In this field, the provision of separate and reliable measures in relation to each of the specific categories of error in punctuation, capitalization, grammar, and usage would demand a test consisting of dozens of parts, each from five to twenty minutes or more in length. The total or composite

THE THEORY OF TEST CONSTRUCTION

score on such a test would doubtless constitute a good measure of general achievement, but not sufficiently better for that purpose than one derived from the much shorter and more conveniently administered mixed-error type of test to justify the added time required.

Markedly different procedures, then, must often be employed in test construction, depending upon which of these two immediate objectives is to be accomplished; each requires a different approach to the whole problem. These facts are extremely significant and should be readily apparent, but in spite of their significance they have received very little attention in the literature to date. It is imperative, therefore, that we begin this discussion of general considerations by discriminating carefully between diagnostic tests and general achievement tests, and that we recognize the problems of test construction that are unique to each.

DISTINGUISHING CHARACTERISTICS OF GENERAL ACHIEVEMENT TESTS AND DIAGNOSTIC TESTS

A general achievement test is one designed to express in terms of a single score a pupil's relative achievement in a given field of achievement. Its principal purpose is to enable us to rank the pupils in a given group in the order of their total achievement within the given field, rather than to measure achievement directly in terms of absolute units or in relation to an absolute standard, as against a standard of "perfection" or against an arbitrary standard such as a "passing grade." (The practice of attempting to express performance on a test as a per cent of possible or "perfect" performance and of estimating in advance an arbitrary per cent value, such as 70 or 75 per cent, as the "passing grade," is one which has only served to confuse many of the real issues in test construc-

GENERAL CONSIDERATIONS

tion. It is a practice which never did serve its intended purpose, even with the traditional essay examination, and which is clearly inapplicable to the newer objective types of tests. Reasons why this practice should be abandoned will be presented later in this discussion.¹

The term "field of achievement" as used in the preceding definition is, of course, highly indeterminate, its meaning being a matter of arbitrary decision in each special instance. We might, for example, build a test to cover a very broad field, such as general mathematics, general science, or the social studies. Perhaps the majority of general achievement tests are constructed to measure relative achievement in an established school subject, such as physics, chemistry, United States history, plane geometry, English, etc. Many tests, however, are designed to cover only a portion of such fields. For example, we might construct a test for electricity and magnetism in physics, or for Latin grammar or Latin vocabulary in first-year Latin, or for punctuation, capitalization, or spelling in English. Such tests would also be of the general achievement type. The distinguishing characteristic of a general achievement test, then, as it is here considered, is that performance in a given "field," whatever the limits of that field, is expressed in terms of *a single score*, and that the test is not designed to discover specific weaknesses, errors, or gaps in achievement within that field. Many of our so-called "diagnostic" tests, therefore, may be regarded as consisting of a battery of separate tests, each of which is of the general achievement type. A test in English correctness, for example, designed to yield separate part scores in spelling, punctuation, capitalization, grammar, and usage, may be considered as made up of one test of the general achievement type in spelling, one in capitalization, one in punctuation,

¹ See page 35.

etc. Such a test would be broadly diagnostic in the sense that it discriminates between these major categories in English correctness, but it might not be intended to discover specific weaknesses within, for example, the field of grammar. That is, the part concerned with grammar alone would not itself be considered as a diagnostic test. If in this situation the part scores were totaled to yield a single score for English correctness as a measure of general achievement in the whole field, then the test as a whole, as well as each of its separate parts, could be considered as a test of the general achievement type.

A diagnostic test, as has already been implied, is one intended to discover *specific* deficiencies in learning or teaching. It is a test in which a single total or composite score is of little or no significance, but on which the part scores or the percentages of correct responses to individual items are the measures sought. Tests may be diagnostic in various degrees. A test in English correctness, for example, may break the whole field up into such divisions as spelling, capitalization, punctuation, grammar, and usage, yielding a part score for each division, or may still further analyze each of these divisions, splitting the section on punctuation into tests of the use of the comma, period, semicolon, etc. Or it may make an even more detailed analysis, considering separately, for example, each of the types of situations in which, for instance, the comma may be used. To the degree, then, that the emphasis in the test is placed upon the part scores or upon percentages of responses to individual items, that test is of the diagnostic type. To the degree that the emphasis is placed upon a single total score, designed to yield a measure of general achievement, the test is of the general achievement type.

Perhaps the majority of the tests constructed for informal use by the classroom teacher are or should be of the *diagnostic*

GENERAL CONSIDERATIONS

type. This is particularly true of the tests administered during the course of instruction to serve as the basis for remedial teaching. Tests used as "final examinations," however, and most standardized tests, including many of those used in cooperative regional testing programs, are definitely of the general achievement type.

The same test may, of course, be intended to serve both of these immediate objectives, that is, it may at the same time be considered as a test of general achievement in a given field and as an instrument for diagnosing achievement within that field. Many such tests have in fact been constructed. This practice may in large part account for the failure to recognize the legitimate functions and essential characteristics of each type of test, as well as for the failure to recognize that the two types cannot often be effectively combined in a single instrument. Test users have been encouraged to expect both types of service from a single test, and as a result many of the general achievement tests now being constructed have been subjected to widespread but unfair criticism. What is needed, therefore, is a clearer understanding on the part of all concerned of the technical limitations of a test of the general achievement type, and a more adequate appreciation of the precise functions which, in view of these limitations, it may reasonably be expected to perform.

THE NATURE AND FUNCTION OF A GENERAL ACHIEVEMENT TEST

The Significance of the Single Score Feature

The most important of the restrictions placed upon the content of a general achievement test arises from the significance of the single-score feature of such tests and from the necessity for sampling.

THE THEORY OF TEST CONSTRUCTION

It should be noted, first of all, that the meaningfulness of a single score as a measure of achievement in any field depends upon the *homogeneity* of that field. If the specific achievements in a given field are capable of classification into distinct types (perhaps with reference to independent objectives of instruction) between which only very low or negative relationships exist, then no single measure can adequately describe the student's relative status in all of these types simultaneously. For example, if within the "field" of high-school chemistry, "extensiveness of scientific vocabulary" happens to be negatively correlated with "skill in manipulation of laboratory apparatus," so that students superior in one aspect tended to be inferior in the other and *vice versa*, then it would be relatively futile to attempt to measure both of these types of achievement in a single general achievement test yielding only one score. In this instance, a highly meaningful description of the individual's status could be obtained only if each type of achievement were independently measured and considered.

There is, of course, no field of achievement for which general achievement tests are now being constructed that is perfectly homogeneous in the sense indicated, i.e., there is no field in which all of the constituent types of achievement are perfectly correlated. In any field, therefore, the use of a single general achievement test may and often will hide the fact that in certain types of achievement *within* that field an individual has deviated significantly from his own general level. A student may, for example, make a high total score on a general achievement test in English correctness and yet be relatively weak in certain word usages, a fact which could not possibly be indicated by his total score. It certainly does not follow, however, that because of this limitation we must analyze each field into homogeneous or "unitary" skills, traits, or abilities and measure each separately in order to

GENERAL CONSIDERATIONS

get any meaningful result. Rather, we recognize the limitation and accept it in exchange for the greater convenience in administration and interpretation of a single general achievement test as compared to a diagnostic battery of tests of special abilities. Even though we did measure each ability separately, it would still be necessary, for many practical administrative purposes, to combine the measures into a single composite score, or subjectively to base a judgment of a composite type upon them. The general achievement test, then, is in a sense simply a device for securing conveniently, and on a *comparable* basis for all students, a composite description of total achievement in those situations where a composite type of interpretation must in any event be made.

While the use of general achievement tests may therefore be readily justified, we nevertheless must recognize that it has these limitations, and that they become more and more serious as the field becomes less homogeneous. In *all* fields of achievement, as usually defined under the present organization of the curriculum, the general achievement test is definitely restricted in usefulness and, for many significant purposes, must be supplemented by tests or test batteries of a diagnostic character.

In any of the fields for which we now construct general achievement tests, it is obvious that no single test of this type can hold the pupil directly responsible for the development of, understanding of, ability to use, or even verbal learning of all of the specific information, relationships, ideas, generalizations, skills, abilities, and attitudes which constitute that field. The subject matter of United States history, for example, consists of thousands of items of information, and of thousands of related ideas, generalizations, inferences, implications, etc., based upon that information which it might be considered desirable that the pupil learn and understand.

THE THEORY OF TEST CONSTRUCTION

If each of the specific elements of achievement in a given field could be assigned a weight proportional to its importance or value, then the pupil's total achievement could be measured by a single score which would be equal to the weighted sum of the elements that he has actually mastered. This concept of a *true* measure of general achievement, of course, can be only hypothetical. Even for a very restricted field, no single test could be constructed and administered which would measure each of its many elements directly and independently in separate test items.

The items constituting any given general achievement test must therefore be considered as representing only a very restricted *sampling* of all of the items that might be constructed on the basis of the subject matter involved. Few teachers will use a general achievement test that requires more than two hours for administration, and most of them prefer a test of shorter length. Very few tests, even of the objective type, can include much more than from 100 to 200 items. With so restricted a sampling, it is highly important that each element in the sampling or each item in the test contribute as much as possible to the validity of the whole test. The validity of the whole test depends upon the degree to which the single scores obtained from it rank the pupils tested in the order in which they would be ranked by a true measure of total achievement such as was hypothesized in the preceding paragraph. It follows that the validity of any *single item* in the test must also depend (within limits) upon the degree to which that item of itself discriminates between pupils of inferior and superior total achievement.

The Difficulty of General Achievement Test Items

It follows immediately that no item which is answered correctly by *all* pupils in a given group can be of any functional

GENERAL CONSIDERATIONS

value in a general achievement test for that group, nor can any items which are answered correctly by *none* of the pupils. Any item to which all pupils in a group respond alike, whether correctly or incorrectly, obviously cannot serve to *discriminate* between those pupils. Items of this kind, of zero or 100 per cent difficulty, therefore, have no place in a general achievement test, since they contribute nothing to its basic purpose.

This immediately establishes a fundamental difference in the content of general achievement and diagnostic tests. In a diagnostic test it may be very desirable to know that there are certain things which have been learned or certain skills which have been acquired by all or by no pupils in a given group. Such information may be of value in planning remedial instruction or in analyzing the instructional needs of pupils. In diagnostic tests, then, it is legitimate to include any items about which information concerning the achievement of pupils is desired. In tests intended only for the measurement of general achievement, however, only those items should be included which will be answered correctly by some but not all of the pupils in the group to be tested.

From this it follows that frequently some of the most important items of achievement are of least functional value in a general achievement test. This is because, if instruction has been adequate, many of these very important or fundamental items have been so thoroughly taught as to have been mastered by all pupils. Lists of fundamentals or of the minimum essentials in a course of study are, therefore, very often a poor source of material for general achievement test construction. An item testing for the year in which Columbus discovered America, for example, is rarely of any functional value in a general achievement test in high-school American history. The 100 addition facts may prove valid material for a diagnostic test in arithmetic, but most of them are un-

likely to function in a general achievement test of that subject, particularly at the higher levels of instruction.

The Range of Difficulty of Items

To be of any functional value in a general achievement test, each of the items must be missed by some pupils but not by all. It is essential, furthermore, that the items be *distributed* along the difficulty scale. Items of different degrees of difficulty will discriminate between pupils at different levels of ability or achievement. Certain items will be so difficult as to be missed by all pupils who are inferior in general achievement, and hence, such items will be of no value for discriminating between pupils at these lower achievement levels. These difficult items are essential, however, if proper discrimination is to be secured between superior pupils. Other items will be so easy that only the very inferior pupils can fail to respond correctly, and hence will be of value only for discriminating between the poorer students.

A test that is to discriminate properly between pupils at all levels of achievement in a given group must contain therefore items of all degrees of difficulty for that group. If it contains nothing but easy items, the scores of the better pupils will be heavily concentrated at the top of the scale, and there will be many ties at or near the highest possible score. If the test contains nothing but difficult items, there will be many ties at or near a score of zero for the poorer pupils. ✓ An examination that is properly adjusted in difficulty to the group tested will result in a "spread" of scores, and in discrimination at all levels of achievement. ✓

Because of this necessity for a range of difficulty in the items, it may not be possible to make the content of the test a representative or random sample of the content of the course of study. Certainly, it will not be possible to limit the content.

of the test to the essential or most important elements in the course. In order to discriminate between pupils at the higher levels of achievement it may be necessary, in some items, to test for a depth of understanding or for a degree of skill that is definitely beyond the average or typical pupil, and that he is not even expected to attain. *It is not implied, however, that one may go outside the course of study to find such items, or that one may draw on trivial or non-essential materials.* ✓ In every properly constructed course of study there is much that has been included for the definitely superior pupil, often with no hope that it may be attained by all. ✓ In every course of study, therefore, there should be sufficient material for the construction of the more difficult items needed. It is no valid condemnation of a general achievement test, then, to find that it contains some items which one does not expect all pupils, or even the typical pupils, to have learned. ✓

The Form of the Distribution of Item Difficulty

Test authorities are not in agreement upon what is the best *form* of distribution of item difficulty in a general achievement test. Some would favor a rectangular distribution, i.e., about an equal number or proportion of items at each difficulty level. Others propose including relatively few very easy or very difficult items, with the majority near the 50 per cent difficulty level. In general, however, they are agreed that there should be a range in difficulty from about 5 to 20 per cent to 80 to 95 per cent, and that the average difficulty of all items should be about 50 per cent. In other words, they agree that the average score made on the test should be about half of the highest possible score. The typical classroom teacher does not usually have the facilities to try out in advance and thus to secure the proper distribution of difficulty in the finished test by selecting from a stock of items of previously determined

difficulty. It is therefore sufficient for the classroom teacher to know that he should try to avoid items that are likely to be missed by all or by no pupils, and should try to make the test as a whole so difficult that the score of the typical pupil will be about half of the possible score. He should also try to adjust the difficulty of the test so that the highest score made will be near but not at the highest possible score, and the lowest score made will be near but not at zero. In other words, the full range of possible scores should be actually utilized if the test is to have the maximum discriminatory value. ✓Any general achievement test should begin measuring at or below the point represented by the pupil whose achievement is lowest, and its scale should extend to a point sufficiently high to include the superior pupil when he is performing at his best. ✓To extend the test scale beyond these limits in either direction is only to add something that will not in any event be used. ✓The "spread" of scores on a general achievement test *should extend from near zero to near the highest possible score.* ✓

These statements apply equally well to both the traditional essay and the newer objective types of general achievement tests, although not, of course, to diagnostic tests of either type.

It may be well to note again that, because of these difficulty considerations, many items may have to be eliminated from a general achievement test that are of fundamental importance in the course of study, and that a relatively large proportion of items must be included which are based upon those materials that have not been well learned by all pupils. ✓Consequently, the content of a general achievement test cannot be expected to parallel exactly the course of study or lists of minimum requirements, and it is distinctly unfair to criticize such tests for a failure to do so if such failure results from an attempt to secure the proper distribution of difficulty in the test items.

Again it should be emphasized that there is no justification for including trivial or irrelevant material to secure the desired range of difficulty.

There may be special instances in which the suggestions made in this section will require some modification. For example, if the same test is intended for administration to a group of pupils both at the beginning and at the end of a course of instruction, in order to secure a measure of progress, then the test should be adjusted in difficulty so that the spread of scores *from both testings combined* will extend from near zero to near the highest possible score. In that case, the upper part of the range of possible scores may not be utilized in the first testing, and there may not be many scores in the lower part of the range in the second testing.

Further modification of the preceding suggestions is required in relation to recognition tests of the alternate or multiple-choice types. In a true-false test, for example, few if any pupils will fail to respond correctly to a significant proportion of the items, since theoretically any pupil will, by chance, select the correct response to approximately half of the items attempted. In such tests, therefore, it would not be practicable to attempt to distribute the items around an average difficulty of 50 per cent, although the scores, if corrected for chance, may range from near zero to near a perfect score. Test authorities do not appear to have committed themselves concerning the optimum range of scores on such tests. The writer offers the opinion, unsupported by statistical evidence, that tests of this character will be most valid when the range of scores, corrected for chance, extends from slightly above zero to slightly below a perfect score. This would mean, of course, that the average percentage of correct responses to individual items will be above 50 per cent by an amount dependent upon the number of choices allowed.

The "Passing Grade"

The concept of the "passing grade," expressed as a per cent of possible or "perfect" performance on a test, is one of the most unfortunate of our inheritances from traditional examination practices. This concept has served only to confuse many of the real issues in general achievement test construction. Because of this concept, for example, many of the statements made in the preceding paragraphs, particularly in so far as they apply to essay or discussion types of examinations, will appear difficult for many teachers to accept. The statement that the average score on a general achievement test should be about half of the highest possible score cannot be reconciled, of course, with the practice of setting the passing grade at 60, 70, or 75 per cent of the highest possible score. If, therefore, we are to accept the suggestions made concerning the proper distribution of item difficulty, we must be prepared to discard the concept of the "passing grade."

The position of a "passing grade" along a scale of possible scores on a general achievement test is *completely irrelevant* to the task of general achievement test construction. The problem for the test constructor is to discriminate between pupils in what they actually have achieved, regardless of how what they *have* achieved relates to what they presumably should have achieved. The position of the passing grade must always be arbitrarily and independently determined, not in terms of the *per cent* of items answered correctly in a given test, but in terms of a description of the detailed items of achievement which collectively represent the minimum of "satisfactory" achievement in the subject or course involved. Whether 10, 70, or 90 per cent of the pupils exceed that score has no bearing whatsoever on the effectiveness of the test for *ranking* pupils in the order of their actual total achievement.]

GENERAL CONSIDERATIONS

If there were available for a given course a detailed description of the specific elements of achievement which individually and collectively are considered *essential* to promotion, and if each, or if a random sample, of these elements could be measured by separate test items, then a "passing grade" could be meaningfully described, but in this case it would be *at or near 100 per cent*, and not at 60, 70, or 75 per cent. That is, a pupil would deserve promotion only if he had mastered *all* of the minimum essentials for promotion. Such a "mastery" test, however, would be of little or no value for discriminating between pupils at different levels of achievement above the minimum, since presumably it would result in perfect or near perfect scores for most of the pupils. In other words, such a "mastery" test would not be considered a general achievement test as here defined. While such tests might be desirable, there are few if any high school or college courses for which the "minimum essentials" have been authoritatively described in a form sufficiently specific to make possible the construction of such tests. For this reason alone, if for no other, the idea of an absolute standard and of grades or scores with absolute meaning must be abandoned. Other considerations point to the same conclusions.

Presumably, when a student earns a grade of 70 per cent on an examination, he has attained 70 per cent of "perfect" performance on that test, or has learned or mastered 70 per cent of that which has been given him to learn. This, of course, is rarely, perhaps never, the true meaning of such a score. There is no definite limit to achievement in any course, nor is there any definite meaning to "perfect" performance on any essay test. The 100 per cent standard, therefore, has no single meaning, but varies from test to test in a manner which is largely accidental, since the difficulty of a test cannot be accurately anticipated or controlled by the

classroom teacher under the conditions of informal test construction. What actually happens in using this system is that the teacher consciously or unconsciously adjusts the test and the scoring of it so that a predetermined number of students will "pass," regardless of the per cent value at which the passing grade is set. If the teacher constructs a test and finds, upon administration, that it proves more difficult than he anticipated, he becomes more lenient in his grading so that too many students will not be below passing. If it is easier than he anticipated, he again adjusts his standards so that a few of the poorest students will "fail." (Unfortunately, this revision in standards is often made after a number of the papers have been graded, with the result that the first and last papers are not graded on the same basis and hence do not receive comparable grades.)

The rigidity or laxity of the "standards" employed, furthermore, is in no way indicated by the point at which the passing grade is set. If the passing grade is set at 60 per cent, the teacher will adjust the difficulty of his test or the method of scoring it so that what he considers the "proper" proportion of students will exceed that score. If it were set at 75 per cent, he would see to it that the same number of pupils would "pass." Under this system, such results are inevitable. In the absence of any authoritative, specific description of minimum essentials, the teacher must decide subjectively, in terms of the students' relative (not absolute) test performance and on other bases, which students should "fail," and then so grade the papers that these students will receive a mark below "passing," regardless of the per cent value at which the passing mark has been set.

In assigning per cent grades with reference to an arbitrary "standard" in any test, we are in large measure only deluding ourselves to no particular advantage. The per cent system

GENERAL CONSIDERATIONS

appears to set absolute standards and to result in scores which are comparable from test to test, but in reality it does not. Scores on general achievement tests, whether of the essay or the objective type, can only have relative meaning, and the system of scoring should be consistent with this truth. The per cent system, with its arbitrary passing grade, should be completely discarded, and an admittedly relative system, such as the point scoring system, should be substituted for it.

This, of course, is what is done with tests of the objective type, in which no attempt is made to express scores in terms of per cents, and in which no attempt is made to keep the number of items, or the highest possible score, at 100 or at any other fixed number. In tests of the essay type, the consistent and preferable procedure would be to assign an arbitrary number of maximum points or credits for each question, in proportion to the relative importance of the questions if such a basis is considered desirable, without any regard to what may be the sum of such possible points for the entire test. Points should then be assigned to each answer on the basis of its *relative* quality, the best answer being given the maximum number of points and the poorest answer being given no credit.

Whether or not such a reform will ever be accomplished in the scoring of essay tests is a question of minor concern for the purposes of the present discussion. The foregoing considerations have been presented, not so much in the hope that a practice so deeply rooted and steeped in tradition could be changed by rational argument, as to support the statements made earlier with reference to tests of the *objective* type. It has perhaps been primarily because of the "passing grade" concept that teachers have been slow to accept the principle that the difficulty of an objective test should be so adjusted as to secure a "spread" of scores from near zero to near the

possible score, with the *average* score about half of the possible score. Because of its bearing on this principle, this discussion of the "passing grade" has been considered essential.

THE VALIDITY OF A SINGLE TEST ITEM

It has already been noted that, if an item is to *function* in a general achievement test, it must *of itself* discriminate between pupils at different levels of general achievement in the field tested. In order to do so it must, among other things, be missed by some but not all of the pupils tested. This characteristic, however, is not alone sufficient.

It is a common misconception that if a test item actually holds the pupil responsible for an element of information, a skill, an ability, an understanding, an insight, or for any other trait which unquestionably belongs to and has an important place in the subject tested, and if it is neither so difficult that all pupils fail nor so easy that all pupils succeed, the item is for these reasons valid for inclusion in a general achievement test. In other words, it is often wrongly believed that an item may be judged valid for inclusion in a test solely upon the basis of its difficulty and upon its desirability for inclusion in the course of study. This misconception again arises directly out of the failure to discriminate between the functions of diagnosis and of the measurement of general achievement, and out of a mistaken notion that these functions can be performed effectively by a single instrument.

It often happens that two objective test items may prove equally "difficult" and may hold the pupils responsible for equally valid content from the curriculum viewpoint, and yet the actual responses made to one may be much more highly related to general achievement than those made to the other. Certain items, apart from their difficulty or desirability, repre-

GENERAL CONSIDERATIONS

sent far more crucial tests or indicators of a pupil's level of *general* achievement than others. It is obvious, for example, that if a plane geometry test item were accidentally included in a test in Latin, that item might be answered correctly by about 50 per cent of the Latin pupils and hence would be acceptable as far as its difficulty is concerned; yet it would have very little value for the purpose of discriminating between pupils who are superior and inferior in their general achievement *in Latin*. Were an item from English grammar accidentally included in the Latin test, it also might have an acceptable difficulty, and might even, because of the higher relationship between ability in English and Latin grammar, be of some value in distinguishing good Latin students from poor ones, although its discriminatory power in Latin would not be expected to be high. Such items, of course, could be easily detected and eliminated from the Latin test because they obviously do not "belong" to Latin. Among the items that clearly do *belong* to the field of Latin, however, there may be many that have no more discriminating value in that field than did the geometry and English items. The mere fact that an item "belongs to" a given field is no guaranty that it will discriminate effectively between pupils at different levels of achievement in that field.

✓The worth or effectiveness of a test item depends, therefore, not only upon its desirability for inclusion in the curriculum and upon its "difficulty," but also upon its power to discriminate between pupils of high and low levels of general achievement in the field involved. ✓It is important that we recognize this double aspect of the validity of a test item, and that we see clearly the relation between "validity" from the curriculum viewpoint and "validity" for achievement testing purposes as determined by the discriminating power of an item. The nature of this relationship and its

various implications are discussed briefly in the following section.

Much of this discussion will be mainly of only theoretical interest to the classroom teacher, who has not the facilities for determining objectively, through preliminary try-outs, the discriminating power of individual test items. It will, however, make clear the reasons for some of the specific suggestions for test construction that will later be given, and will provide the reader with a more adequate technical vocabulary for the succeeding discussions.

Before proceeding, it may be well to remind the reader again that we are here concerned only with general achievement tests. The considerations here presented should not be taken to apply to items in a diagnostic test, on which the total score is of no significance. They are important in general achievement testing only because of the restricted sampling employed, and would be of little significance there if a sufficiently broad sampling could be taken to insure high validity.

The Discriminating Power of a Single Test Item

The discriminating power of a single test item refers to the degree to which success or failure on that item by itself indicates possession of the ability which is being measured. In relation to tests of the general achievement type, it may be defined as the accuracy with which a pupil can be placed along the scale of general achievement on the basis of his success or failure on the given item. An item may be said to have perfect discriminating power if every pupil who responds correctly to the item ranks higher on the general achievement scale than any pupil who fails on the item. An item may be said to have zero discriminating power when there is no sys-

GENERAL CONSIDERATIONS

tematic difference between the general achievement of the pupils who succeed on the item and those who fail.

Otherwise stated, an item is said to discriminate if the pupils who respond correctly to that item are, on the average, superior in *general* achievement to those who respond incorrectly. If the pupils who succeed on a given item are, on the average, just equal in general achievement to those who fail, then the item has no discriminating power. The degree of discriminating power of an item, therefore, depends upon the magnitude of the difference between the *averages* in general achievement of those who succeed and those who fail on the item.

One method of determining the discriminating power of a given test item, therefore, would be to administer a comprehensive criterion test of general achievement to a large group of pupils and then to compute the average score on the criterion test for the students who succeeded on the given item and for those who failed on the item. If the average score on the criterion test of those who succeeded on the given item were higher than that of those who failed on the item, then that item could be said to have discriminating power, the degree of discriminating power which it possesses depending upon the magnitude of the difference in the two averages.

Another method would be to break the total group tested into a number of sub-groups according to their scores on the criterion test and then to compare percentages of correct responses to the given item for the pupils in the various sub-groups. Suppose, for example, that a criterion test of general achievement in a given field is administered to a group of 1000 pupils and that this group is split into ten sub-groups as follows: group A consists of pupils scoring below the 10th percentile on the criterion test, group B of those scoring between the 10th and 20th percentiles, group C between the 20th and

THE THEORY OF TEST CONSTRUCTION

30th... and group J the highest or those scoring above the 90th percentile. Now if a given individual test item is to possess discriminating power in the field involved for the total group tested, it must show a relatively low percentage of correct responses for sub-group A, a larger percentage of correct responses for sub-group B, and an increasing percentage of correct responses for each of the succeeding sub-groups. The facts for a single item might be graphically represented as follows:

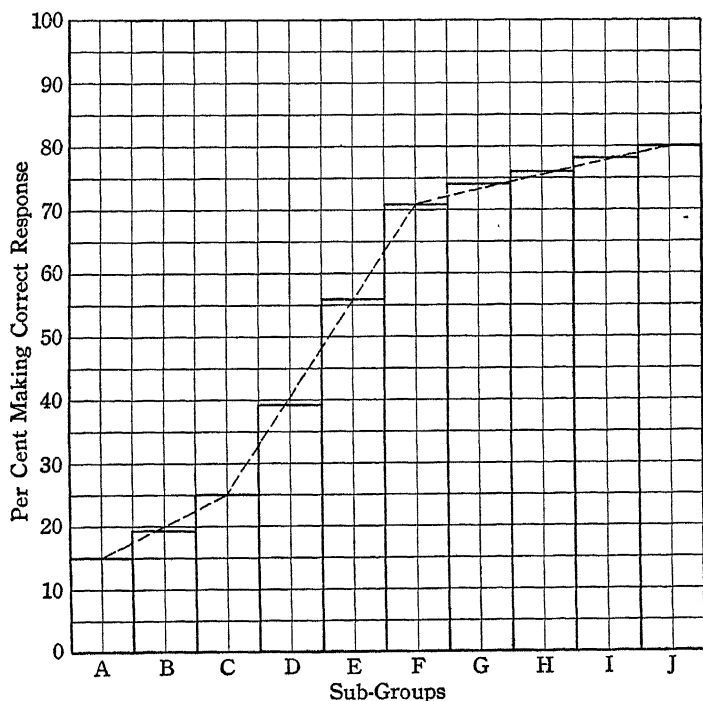


FIGURE 1. PERCENTAGE OF CORRECT RESPONSES TO A GIVEN TEST ITEM BY PUPILS AT VARIOUS LEVELS OF TOTAL ACHIEVEMENT

GENERAL CONSIDERATIONS

For the item represented in this figure, only 15 per cent of the pupils in sub-group A made the correct response, as compared to 80 per cent in sub-group J. (The height of each rectangle represents the per cent of the pupils in a given sub-group who responded correctly to the given item.) The same facts could be represented by a single broken line, such as that joining the midpoints of the upper bases of the rectangles in Figure 1, which could be described as the "line of discrimination" for the item. The sharper the rise in this line, the more highly discriminating will be the item represented. An item showing a straight horizontal line with no rise or fall would be an item of zero discriminating power, and one showing a line falling from left to right, rather than rising, would be a negatively discriminating item.

Various hypothetical degrees of discriminating power for a test item of 50 per cent difficulty are represented in Figure 2. This figure shows the various types of relationships which may be found between general achievement, as measured by a comprehensive criterion test, and the ability to respond correctly to a single given item, in this case an item answered correctly by 50 per cent of the pupils in an experimental group. The vertical scale in this figure indicates the per cent of pupils who responded correctly to the item. The placement of pupils along the general achievement scale (the horizontal) is determined on the basis of their percentile standing on the criterion test. The "line of discrimination" for a given item indicates the percentage of pupils at each level of general achievement who responded correctly to the item, and has been constructed after the manner illustrated in Figure 1. The lines in Figure 2, however, have been smoothed, rather than shown in their original broken form.

Line *MM* represents the line of discrimination for an item of 50 per cent difficulty which shows perfect discriminating

THE THEORY OF TEST CONSTRUCTION

power, since every pupil below the 50th percentile of general achievement missed the item, and every pupil above the 50th percentile succeeded on it. The pupil who responds correctly to this item may be accurately placed on the general achievement scale with reference to one point, in this case the point of median achievement. Only a dichotomous classification, of course, is possible; the pupil's distance above or below the median point in general achievement cannot be determined on the basis of this item alone.

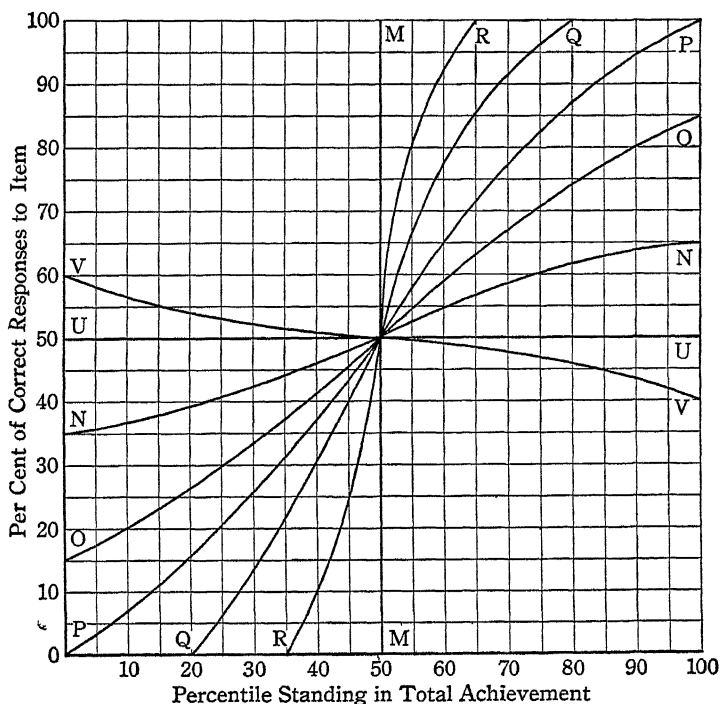


FIGURE 2. HYPOTHETICAL LINES OF DISCRIMINATION SHOWING STAGES BETWEEN PERFECT AND MINUS DISCRIMINATION POWER FOR ITEMS OF FIFTY PER CENT DIFFICULTY

GENERAL CONSIDERATIONS

Line *UU* represents the line of discrimination for an item of 50 per cent difficulty which has zero discriminating power, since the same per cent of pupils at every achievement level responded correctly to the item. When a pupil responds correctly to an item of this type, there is no greater reason for placing him on the lower part of the general achievement scale than on the upper, i.e., he is no more likely to be good than poor in general achievement. Items of this type have no functional value in a general achievement test, regardless of their other characteristics.

Line *VV* represents the line of discrimination for an item of 50 per cent difficulty which has negative discriminating power, since it was answered correctly more frequently by pupils of inferior general achievement than by pupils of superior achievement. A pupil who responds correctly to an item of this type is more likely to be *low* in general achievement than one who responds incorrectly. Such items can have no functional value in a general achievement test, unless the pupils are given credit for making the *wrong* response, which obviously is impracticable. A large number of items of this type have been discovered by the author in experimental try-outs of test materials in the high-school subjects, and concrete illustrations of them will be presented later.

Between the extremes of perfect positive and negative discriminating power, all degrees of discrimination may be found. These are illustrated for items of 50 per cent difficulty only, in Figure 2 by lines *NN*, *OO*, *PP*, *QQ*, etc.

It is very important to note that there is no apparent reason for assuming any relationship between the discriminating power of a test item and its "difficulty," or percentage of incorrect responses. For example, while only 100 out of 1000 pupils may succeed on a given test item, it may happen that these 100 pupils are on the average no higher in *general*

THE THEORY OF TEST CONSTRUCTION

achievement than the 900 who fail on the item in question. Similarly, an item may be answered correctly by 800 out of 1000 pupils, i.e., it may be an "easy" item, and yet among the 200 pupils who failed on this specific item there may be many who are superior in general achievement to some of those who succeeded on it. The difference in average achievement between those pupils who fail and those who succeed may be much greater for one item than for another of the same difficulty, or greater for a given "easy" item than for another

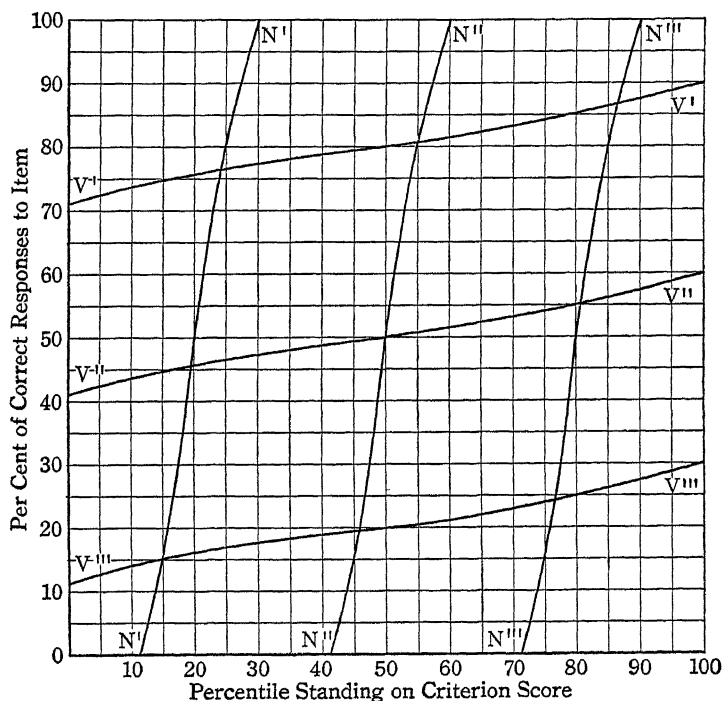


FIGURE 3. HYPOTHETICAL LINES OF DISCRIMINATION ILLUSTRATING HOW ITEMS OF VARIOUS DIFFICULTIES MAY BE EITHER HIGH OR LOW IN DISCRIMINATING POWER

GENERAL CONSIDERATIONS

that is very difficult. An item of any difficulty may have any degree of discriminating power.

In Figure 3, hypothetical lines of discrimination are illustrated for good and poor items at each of three levels of difficulty.

Line $N'N'$ represents an item of 80 per cent correct responses (20 per cent difficulty) with very high discriminating power, and line $V'V'$ represents an item of the same difficulty but with very low discriminating power. The lines $N''N''$ and $V''V''$ represent items of 50 per cent difficulty with high and low discriminating power respectively. Lines $N'N'$, $N''N''$, and $N'''N'''$ represent items of marked differences in difficulty but with the same degrees of discriminating power. Pupils may not always be assumed to be high in general achievement simply because they succeed on a "difficult" item, or low in achievement simply because they fail on an "easy" item. There are many "difficult" items which are more frequently answered correctly by inferior than by superior pupils, and many "easy" items which are more frequently missed by good than by poor pupils.

The degree of discriminating power of an item therefore depends upon the sharpness of the rise in its line of discrimination. A sharp rise in a line of discrimination, however, may occur at either the lower or the upper end of the ability scale. The item represented by line $N'N'$ in Figure 3 is just as discriminating as that represented by the line $N'''N'''$, but the first of these items discriminates between pupils of inferior ability, while the latter discriminates between pupils of superior ability. The first item would be considered as relatively easy in terms of the percentage of correct responses for the total group, while the latter would be considered relatively difficult; but both would be equally discriminating at their respective levels.

In terms of this illustration, it should be apparent that the ideal test would consist of items of high discriminating power distributed evenly over the difficulty scale. In other words, the ideal test would contain some easy items that discriminated sharply at low levels of achievement, others of medium difficulty that discriminated sharply at intermediate levels, and still other difficult items that discriminated sharply at high levels of achievement. The significance of this ideal distribution of item difficulty has already been partly discussed, and will be considered again later.

Statistical techniques have been developed for expressing the discriminating power of a test item in terms of a single numerical index, or "index of discrimination." Because of the technicalities involved, no attempt will be made here to describe the specific nature of these indices. In the later discussion, however, the numerical values of one of these indices of discrimination will be quoted for illustrative test items.² It will be sufficient for present purposes to explain that the index used has the following characteristics: its value will be unity (1.00) for items of perfect discriminating power, i.e., for items whose line of discrimination shows an abrupt vertical rise at any point along the scale; its value will be zero for items of no discriminating power, i.e., for items whose line of discrimination will be a straight horizontal line; its value will be negative for items of minus discriminating power, i.e., for items whose line of discrimination will fall rather than rise in going from left to right along the ability scale; its positive value will be in direct proportion to the sharpness of the rise in the line of discrimination, i.e., an item showing a sharply

² The index here used is the bi-serial coefficient of correlation between responses (right or wrong) to the given item and the scores on the criterion test. For a discussion of this and other indices, see "Experimental Techniques in Test Evaluation," by E. F. Lindquist and W. W. Cook, *Journal of Experimental Education*, March, 1933.

GENERAL CONSIDERATIONS

rising line will have an index with a value near unity, while an item with a line of discrimination showing a slow or gradual rise will have an index whose value is near zero; and finally, there is no relationship between the difficulty of an item and its index of discrimination, i.e., an easy item will tend to have neither a lower nor a higher index of discrimination than will a difficult item.

Practical Limitations of the Index of Discrimination in Evaluating Test Items

The power to discriminate between pupils at various levels of general achievement is theoretically an important characteristic of the items in a general achievement test. As was noted in the preceding paragraph, various statistical techniques have been devised for the measurement of this characteristic. These techniques are of little practical value for the construction of informal classroom examinations, since the classroom teacher does not have the facilities for preliminary try-out of materials with relatively large groups of pupils, but they may be used to advantage by the constructor of a refined and formal standardized test. There are, however, a number of practical limitations to their use which are of considerable significance.

In the first place, the use of any of these indices of "goodness," "selectivity," or "discriminating power" requires the availability of an independent criterion measure of general achievement, the validity of which may be assumed to be high in comparison to that of the scores on the test whose items are to be evaluated. For example, in order to secure indices of discrimination for the items in a general achievement test in United States history, it would be necessary to have some independent measure of general achievement in that field — a measure whose validity should be very high and at least de-

monstrably superior to that of the scores in the given test. Unfortunately, there are very few fields in which independent criterion measures of this quality can be secured. Frequently the test to be evaluated is itself the best instrument for the measurement of general achievement that we know how to devise for the field in question. If the criterion measure employed is not valid — for example, if it is measuring abilities which are irrelevant to or unimportant in the field in question — then the indices of discrimination computed with reference to this criterion will simply rank the test items in the order in which they distinguish between pupils who are superior and those who are inferior in these irrelevant or unimportant abilities. Regardless of what criterion is employed, the elimination of items showing negative or near zero indices of discrimination with reference to this criterion will improve the correlation between the criterion and the total score on the test. An increased correlation with the criterion would be clearly desirable, however, only if the criterion itself were of acceptable validity.

The value of an index of discrimination in any particular situation obviously depends primarily upon the validity of the criterion measure against which the individual test items are evaluated. Because of the difficulty just noted, of securing acceptable criteria, it has become the practice in some fields to use the total score on *the test itself* as the criterion against which the constituent items are evaluated. When used in this way, the index of discrimination really becomes a measure of the extent to which the individual items contribute to the *reliability*, rather than to the validity, of the entire test. An index so obtained only enables one to select those items in a test that are individually most effective in measuring whatever the test as a whole happens to be measuring. If the test as a whole is not valid, if it happens to be measuring

GENERAL CONSIDERATIONS

the wrong thing, then the selection or elimination of items on this basis would only make the revised test a more reliable measure of that wrong thing, i.e., the test would become more reliable as it became less valid. The elimination of items showing relatively low indices on this basis may have the effect of securing higher and higher reliability by narrowing more and more the mental function tested. An index of discrimination, therefore, should be used with particular caution for evaluating items in a general achievement test where the criterion employed is the total score on the test itself. It may be of considerable value, in such situations, in identifying items that contain structural or technical imperfections, but it is a dangerous basis upon which to eliminate an item if the content of that item is acceptable in terms of other logical considerations.

Regardless of the criterion employed, the indices of discrimination computed for the items of a test on the basis of the performance of a given group of pupils will determine the relative effectiveness of the items *only for that group* of pupils or for other similar groups. The same item in a test may show a low index of discrimination for one group of pupils and a high index for another, depending upon the nature of their instructional background. Indices of discrimination computed for items on the basis of test results in a single school or in a small number of schools, therefore, may be a very unreliable basis for determining the relative effectiveness of the items when used with pupils in other schools. If a general achievement test intended for widespread use is to be revised on the basis of indices of discrimination computed in a preliminary try-out, the sampling of pupils employed should be selected from a large number and variety of local teaching situations.

Finally, it should be noted that the numerical value of an

index of discrimination is not directly proportional to the desirability of an item for inclusion in a general achievement test. The fact that a given item shows a relatively low index in comparison with other items in the same test does not necessarily mean that it is less essential for inclusion in the test than any of the other items. If, as has already been noted, the field tested includes a type of achievement that is not highly related to other achievements within the same field, then items testing for that achievement will almost invariably show low indices of discrimination. Such a type of achievement may, nevertheless, be as essential a part of the whole field as any other. For example, a test in English correctness including items measuring spelling, punctuation, capitalization, and grammatical abilities may show consistently lower indices for the spelling items than for the other items in the test. This fact may only show, however, that spelling ability is not highly related to the other abilities included in this field, and is no indication that the spelling items are not good as spelling items.

In order to be valid, an English correctness test must contain a due proportion of spelling items, regardless of the relative numerical value of their indices of discrimination, particularly when computed with reference to the total score. If, however, the special ability involved is negatively correlated with the whole field, or shows an extremely low positive correlation, it may be necessary to segregate these items in a separate test in order to get a meaningful description of achievement in the whole field.

One of the principal limitations, then, of the indices of discrimination usually employed in test evaluation is that they are concerned only with the correlations between a single criterion and the responses to individual items, and do not take into consideration the inter-correlations between the

item responses. Theoretically, the best test is that in which the individual items correlate highly with the criterion but show relatively low inter-correlations. If a pair of items shows very high inter-correlations, it may be assumed that they are measuring essentially the same specific ability, and hence the inclusion of both in the same test would involve needless duplication. Items A , B , and C , for example, may show correlations with the criterion (X) of $r_{AX} = .70$, $r_{BX} = .68$, and $r_{CX} = .47$, and their inter-correlations may be $r_{AB} = .95$, $r_{AC} = .36$, and $r_{BC} = .40$. In this case, if item A is selected because of its high index, then item C might be better for inclusion with it than item B , in spite of the lower index for C , because A is already measuring that which is measured by B . It is not safe, therefore, to select items from an experimental edition for inclusion in a final test solely on the basis of the indices of discrimination of the items. In general, the best procedure seems to be to use the index of discrimination solely as a means of identifying seriously defective items, i.e., those with negative or very low indices, but not to employ it in choosing between items showing reasonably high positive indices — the choice in the latter case to be based primarily upon logical rather than statistical considerations.

Factors Influencing Discriminating Power

It has already been stated that a test item whose content is highly desirable for inclusion in the course of study will not necessarily prove to be high in discriminating power. The reasons for this lack of correspondence are many, but they may conveniently be classified as: (1) technical weaknesses in the structure of the item, as, for example, ambiguities, irrelevant clues or cues to the correct response, unfairly misleading elements as in "catch" questions, etc.; (2) insufficient

learning; (3) wrong learning and negative transfer from other learning situations; and (4) lack of homogeneity in the field to be measured. The last of these reasons has been briefly discussed in the later part of the preceding section; the others will receive more extended discussion in the pages to follow.

To illustrate the first of these reasons for the failure of a test item to function, we may note that, because of ambiguity in phrasing, an item may be more frequently missed by superior than by inferior pupils. Suppose, for instance, that a true-false statement is open to two interpretations, and that it is true if interpreted according to the obvious meaning, which may be that intended by the author of the test, and false if interpreted according to the unintentionally included hidden or second meaning. The obvious meaning is that to which the inferior pupils are most likely to respond, and a large proportion of them may therefore respond correctly, depending on the difficulty of the item. The superior pupils, however, may be capable of recognizing that the statement may be interpreted otherwise, and, in the belief that it is the second meaning for which the test author is holding them responsible, they may mark the statement *false*. In consequence, a large proportion of the superior pupils might fail on the item as it is scored, even though all of them knew the correct response to the obvious meaning. An item of this kind, then, may show a negative index of discrimination, even though the concept which the author intended to test may be perfectly valid in the curriculum sense. In other words, the item may prove ineffective because its apparent and intended content did not correspond to its actual or functioning content.

Technical defects in item structure which escape the notice of the author of the test are perhaps the principal reason why many apparently good items fail to function. Extended consideration will be given to such defects, and to the ways of

avoiding them, in the next chapter.³ For present purposes, it is sufficient to note that all defects of this nature are avoidable and can be eliminated if sufficient care is taken in test construction and if adequate opportunity for preliminary try-out of the items is provided. Consequently they have little direct bearing on the essential relationship between validity from the curriculum viewpoint and validity for test purposes, and need not be considered further at this point.

The Effect of Insufficient Learning or Understanding

The failure of an item to function because of insufficient or wrong learning is something beyond the control of the test constructor and may necessitate the elimination of certain items even though they are perfectly acceptable from the curriculum viewpoint and are free from technical imperfections.

Items that fail to function because of insufficient learning or understanding are particularly likely to be found in tests of the recognition type, such as multiple-choice or "best-answer" tests and matching exercises. The multiple-choice exercise consists essentially of a direct question followed by a number of possible responses to that question, only one of which is correct or significantly better than any of the others. In the construction of items of this type, a deliberate attempt is made to make the incorrect responses appear as plausible as possible, in order that the pupil who is uninformed or who does not have a thorough understanding of the point raised may be likely to select an incorrect response rather than to choose the correct response on the basis of superficial rote learning, or by chance, or by random guessing. Consider the following illustration:

³ See pp. 107 ff.

THE THEORY OF TEST CONSTRUCTION

What was the principal reason for the increase of the English-speaking population in the American colonies during the eighteenth century?

- (1) Persecution of the Huguenots (6%)
 - (2) Strife in England between Parliament and the King (25%)
 - (3) The large families being reared by the settlers (12%)
 - (4) Serious religious struggles in England at the time (53%)
- (4 per cent of the pupils omitted the item)

This item was included in a test administered to more than 8000 United States history pupils in Iowa high schools. The numbers in parentheses indicate the percentage of all pupils selecting each response.

This item may be more than a test of the pupil's knowledge that large families were being reared by the American colonists during the period indicated, and of his appreciation of the significance of this fact for the increase of population. Given other information, the pupil should be able to arrive at this correct response by elimination, even though he had not previously learned the correct response. For example, aside from the question raised, the pupil should know that during the eighteenth century England was quite free from serious struggles of a religious character. Enough English history is taught in United States history textbooks to provide him with this information. The fact that 53 per cent of the pupils selected response 4 is some indication that they were not well acquainted with happenings in the mother country during the eighteenth century; it is evidence, furthermore, that the selection was deliberate in most cases, since random "guessing" could never have accounted for so large a proportion. Similarly, the well-informed pupil should have been able to eliminate responses 1 and 2 on the basis of the knowledge that serious political difference between Parliament and the king had been settled before

GENERAL CONSIDERATIONS

the beginning of the eighteenth century, and that the Huguenots were not being persecuted in England during that period and had not been persecuted there at any other time. An analysis of the responses to this item indicates that most of the pupils did not know with any certainty that response 3 was correct, and the item therefore became largely a test of the pupil's understanding of the related information required for the elimination of the incorrect responses.

✓ An item of the multiple-choice type, then, may represent more than merely a test of the pupil's knowledge of the single point raised by the question or of his ability to recall a memorized pat answer to that question. It may also, often to a significant degree, be a test of his ability to relate other information to that directly called for, and of his ability to discriminate among several alternatives on the basis of such related information. It may test for knowledge of what is not correct, as well as for knowledge of what is correct. It may test the certainty and the meaningfulness of the pupil's knowledge, rather than merely his verbal memory of a fact which may have been learned by rote. ✓ A single multiple-choice test item may, therefore, hold the pupil responsible simultaneously for a number of related facts or understandings rather than for only one.

These characteristics may appear to be, and in most cases are, desirable in an item to be used in a general achievement test, but they also may often lead to a complete failure of the item to function properly. The superior pupils in the group tested may have just enough information and understanding to arrive at a certain inadequate or wrong but highly plausible solution to a given problem, but not quite enough information and understanding to be able to select the correct response. An analysis of the responses to the illustration just given showed that the pupils who selected the correct response were, on the average, inferior in general achievement to those who

selected incorrect responses, measured by their average scores on a relatively comprehensive examination over the whole field. That is, the item showed a negative index of discrimination. Apparently, the superior pupils knew that political and religious strife in England had at some time been an important cause of migration to the colonies, but did not know that during the eighteenth century England was free from serious struggles of this nature. This is a case, then, where "a little knowledge is a dangerous thing." What the better pupils did know predisposed them to select responses 2 and 4, whereas the inferior pupils, who did not know even that much, probably resorted to "guessing," and by chance hit upon the correct response more frequently than did their more able classmates. Whether or not this explanation is correct in detail, the fact remains that the item did show a negative index of discrimination and that it therefore did detract from the central purpose of the complete examination, which was to discriminate between pupils at various levels of *general* achievement. In other situations, as in a diagnostic test, it would, of course, be valuable to know that the pupils in a given group are not capable of making the discrimination called for in this item, but the item does not have a place in a general achievement test intended for high-school pupils.

The same point may again be illustrated by another item administered to the same group of pupils, in this case an item with a negative index of discrimination of $-.17$.

What was one of the important immediate results of the War of 1812?

- | | |
|--|-------|
| (1) The introduction of a period of intense section- | |
| alism | (39%) |
| (2) The destruction of the United States Bank | (7%) |
| (3) The defeat of the Jeffersonian Party | (7%) |
| (4) The final collapse of the Federalist Party | (43%) |
| (4 per cent omitted the item) | / |

GENERAL CONSIDERATIONS

The correct response to this item is number 4. Nevertheless, the pupils who selected the first and incorrect response were, on the average, superior in general achievement to those who selected the correct response (number 4). Again, the pupils selecting the first and incorrect response apparently did so because of positive but insufficient learning. They knew that a period of intense sectionalism did set in before the middle of the century, and therefore chose the first response. Apparently they did not know, or failed to recall, that a short period of intense nationalism was an immediate result of the Second War with Great Britain, and that this war, therefore, could not be considered as "introducing" an era of sectional strife. Other pupils, with less knowledge in general, were able to select the correct response since they were not attracted to the first response by a certain knowledge that intense sectionalism did develop in the nineteenth century. (It should be noted, however, that for an abler group of pupils, capable of making the judgment called for, this same item might have shown a high positive index of discrimination.)

There is considerable danger, then, that in the attempt to test for relatively high levels of understanding, the wrong responses may be made so plausible that they will defeat their own purpose. These wrong responses may appeal more strongly to the superior student who has the information that makes them appear plausible than they will to the uninformed student who may not know even enough to recognize their plausibility. The uninformed student, not having the information which makes the wrong response plausible, may avoid it and select the correct response, either by guessing blindly or because he has memorized something which he does not understand, while the better students may be almost invariably misled by their superior but insufficient information

or understanding. The result will be that the poorer students will, on the average, score higher on such items than those who are superior.

In a general achievement test intended for a given group, therefore, an item testing for a high level of understanding will function effectively only if there is in that group a significant proportion of students who have actually attained that high level. If the group tested does not include a reasonable number of such students, the item not only will fail to discriminate as it should but may actually discriminate in the wrong direction.

If this fact were more generally appreciated, there perhaps would be fewer demands for items of the so-called "thought" or "reasoning" variety in a general achievement test intended for widespread use, particularly at the high-school level. There can be no question that the quality of reasoning and the level of understanding actually attained by high-school pupils are, in general, much lower than might be wished. Whether we like it or not, the type of achievement now attained by most high-school pupils is definitely at a low level of understanding and is largely informational in character. Consequently, in the construction of general achievement tests which are to discriminate between pupils at the level at which they are found, it will continue to be necessary to limit a large proportion of the items to well-selected information and to a low level of understanding.

Items of the recognition or multiple-choice type, then, present certain unique problems as compared to items of the recall or essay type. In the essay examination or in the recall type of test an item that is beyond the ability of all pupils tested would simply fail to function at all. Every pupil tested would miss or omit the item, and its presence could therefore be readily detected and the item eliminated in

further testing. Even though it were retained, its presence in the test would have no influence upon the relative magnitude of the scores earned. It might therefore do no positive harm in the test, although it might better be replaced by an item that does function. In recognition types of tests, however, there will rarely, if ever, be any items missed or omitted by all pupils, regardless of the real difficulty of the concept tested. Even in the case of those items based on concepts far beyond the best of the pupils tested, a certain number, even of inferior pupils, may select the correct response by chance or because their rote learning is impervious to the plausibility of the alternate responses, while the superior pupils may consistently select a certain incorrect response which appears plausible in the light of their superior but still insufficient knowledge or understanding. In that case the item will show a negative index of discrimination and will thus detract directly from the validity of the entire test. It is primarily because of the "guessing" element in recognition tests that this problem exists at all.

The Effect of Wrong Learning

Wrong learning, as well as insufficient learning, on the part of the pupils for whom the test is intended may cause an item in that test to show a negative index of discrimination. For example, the following item showed a negative index of discrimination of $-.34$ for a random sample of 1000 pupils selected from more than 8000 pupils in United States history in Iowa high schools. The per cent of all pupils that selected each response to this multiple-choice item is indicated by the numbers in parentheses at the right of the response.

THE THEORY OF TEST CONSTRUCTION

✓In the second half of the fifteenth century the Portuguese were searching for an all-water route to India because

- (1) They wished to rediscover the route traveled by Marco Polo (4%)
 - (2) The Turks had closed the old routes (59%)
 - (3) The Spanish had proved that it was possible to reach the east by sailing westward (10%)
 - (4) An all-water route would make possible greater profits (26%)
- (1 per cent omitted the item)

It will be noted that more than half of the pupils selected response number 2. The negative index of discrimination indicates furthermore that the average achievement of the pupils who selected this response was superior to that of the 26 per cent of the pupils who chose the correct response (number 4). Authoritative historians no longer would accept the second response as a sufficient explanation of Portuguese attempts to round Africa, nor would they deny that response number 4 is the best of those given. An analysis of current textbooks in American history, however, will reveal that these lag behind research and that many of them still present the now disproved explanation: "The Turks closed the old routes." It is not surprising, therefore, that the superior pupils are more likely to select this response than those who have made little or no effective attempt to learn the facts contained in their textbooks. This being the case, the inclusion of this item in the test not only contributed nothing to its effectiveness but even detracted from it. There can be little question, however, that the item is free from technical imperfections or ambiguities, and that it does hold the pupil responsible for an established fact of considerable significance in history.

The following item, also taken from the 1932 Iowa Every-Pupil Test in United States History, similarly showed a negative index of discrimination, in this case of $-.13$.

GENERAL CONSIDERATIONS

America's entry into the World War was largely caused by the

- | | |
|---|-------|
| (1) Fear that the defeat of the Allies would lead to the overthrow of republican government in France | (16%) |
| (2) Violation of Belgian neutrality | (45%) |
| (3) Fear of losses by the moneyed interests if the Allies were defeated | (30%) |
| (4) Declaration of war by Italy | (5%) |

(4 per cent omitted the item)

While this item deals with a controversial interpretation, and while there consequently may be some disagreement about which constitutes the best response, it nevertheless provides an excellent illustration of what may happen when the pupil is held responsible for a controversial opinion. Pupils have been encouraged, both as the result of direct instruction and as the result of popular prejudice or unreasoning patriotism, to believe that altruistic or idealistic motives underlie important national actions or policies. In this case, the 30 per cent selecting response number 3, which was keyed as correct, were in general inferior in total achievement to those who selected the other responses. This is only to be expected, since the better pupil learns what he has been taught and encouraged to believe, and the more widely he reads "popular" history with a nationalistic tinge, the more likely he is to search for altruistic motives in an item of this type. While response number 3 undoubtedly would be accepted by many historians as an underlying motive for America's participation in the World War, certainly it is not widely taught in our schools at present.

It is not uncommon for the pupils in our public schools, because of biased teaching and unreasoning prejudices, to ascribe flattering motives for certain national actions or policies in cases where historians will agree that the real and

fundamental motives were undoubtedly of a different character. Items based on these situations and keyed to correspond to the opinion of authoritative historians rather than to popular opinion will frequently show negative indices of discrimination. Items dealing, for example, with issues such as the justification and responsibility for the Mexican and the Spanish-American Wars, or with the moral aspects of actions prompted by imperialistic ambitions, such as the annexation of Texas and the "purchase" of the Philippines, or with the "right" of the Confederate states to secede, will very frequently constitute poor material for the construction of general achievement tests if keyed to agree with the interpretations of eminent present-day historians. Items that have been wrongly learned will occur most frequently in the social studies, which deal with a large number of controversial issues and with social problems about which there exist many fallacious but popular prejudices and misconceptions. They will also occur, however, in other fields, perhaps particularly in the physical sciences, such as physiology, biology, and general science, as a result of the many popular superstitions and misconceptions concerning natural phenomena. To include items dealing with such misconceptions in recognition tests may frequently lower the validity of the test for the purpose of measuring general achievement, and such items must therefore frequently be eliminated from general achievement tests even though they would be of distinct value in diagnostic testing. Again, this problem is one which is significant only in the construction of general achievement tests, particularly of the objective type, and is one which need not concern the builder of a diagnostic test. If a field contains a significant proportion of items of this character, particularly if the items are themselves homogeneous, it would appear desirable to include them in a separate test for independent consideration.

The Apparent vs. the Functioning Content of Test Items

What a test *appears* to measure, or what its *name implies* that it measures, and what it actually does measure may be and often are far apart. A given test may appear, upon a *logical* analysis of its verbal or thought content, to represent a very well selected sampling of a given field of achievement. Hence, it may appear highly valid from the curriculum or subject-matter point of view, yet a *psychological* analysis of the students' reactions to the individual items may show that much of this content is actually non-functioning, and that the test is seriously lacking in validity. In any type of test exercise, but particularly in objective exercises of the recognition types, such as the true-false statement and the multiple-choice and matching exercises, the basis on which the student tested actually responds to the item in selecting the correct response may often differ quite markedly from that which the test author intended. The specific element in the test item which actually determines the student's response may be an element of whose presence the test author is not even aware, or may be one which is only a minor or unimportant detail of the total content of the item.

Unintentional and undiscovered differences between the apparent or intended and the functioning content of individual items constitute a major weakness in many tests of the objective type. The detection and elimination of such defects is one of the most crucial problems in test construction. The purpose of the discussion at this point is to identify this problem and to demonstrate its significance through a series of concrete illustrations. That is, we are here concerned only with showing that the problem exists and that it is of real significance. More specific suggestions for dealing with this problem will later be made in the discussions of the various

types of objective test exercises and in the chapters dealing with the separate subject-matter fields.

The apparent and the functioning content of an item may differ in a very wide variety of ways and types of situations, only a few of which can be considered here. For present purposes we may classify these situations roughly into two types, as follows:

1. Those in which the student may select the correct response on the basis of a verbal cue or clue which is completely irrelevant to the intended purpose of the item.
2. Those in which the functioning element is only a minor part of the content which was intended to function as a whole, i.e., those in which some or a major share of the apparent content is *non-functioning*.

Irrelevant Cues or Clues and Specific Determiners

1. Clues provided by grammatical structure of the item

Illustration 1⁴

Directions: Read carefully topics A and B that follow this paragraph. Then write A on the blank preceding each statement below that refers to the topic marked A; write B on the blank preceding each statement that refers to the topic marked B; write X on the blank preceding each statement that does not refer to either A or B.

A. Constitution of the United States

B. Weaknesses of the Articles of Confederation

X. Unrelated Statements

- 1. It became known as the "supreme law of the land."
..... 2. Congress had no authority to regulate commerce, and therefore was unable to settle quarrels between the states.
..... 3. It created a strong central government.

⁴ From course of study test in American History, Kansas City, Mo., Public Schools, May, 1931.

GENERAL CONSIDERATIONS

- 4. Government under it went into effect in 1789.
- 5. A national supreme court was established to interpret the laws.
- 6. Congress could declare war, but could not raise or support an army.
- 7. It provided for three departments of government.
- 8. There was no real central control.

In responding to this item, the wide-awake pupil will note immediately that topic A is in the singular, while topics B and X are in the plural. It will therefore be at once apparent to him that items 1, 3, 4, and 7 can only apply to topic A. All the pupil needs to know in order to make the correct response to these items is a sense of grammatical consistency. Furthermore, the pupil is immediately disposed to mark items 2, 6, and 8 as relating to topic B, if for no other reason than that these statements *sound like* statements of weaknesses, even though he may not *know* that they were weaknesses of the Articles of Confederation. Clearly, then, a wide-awake pupil could make a fairly high score on this exercise, even though he knew nothing whatever about what the item was intended to measure. This item might discriminate between pupils who are mentally alert and those who are not, but it would not necessarily discriminate between pupils on the basis of their knowledge of history.

This matching exercise was administered to a group of 443 high-school pupils who had previously taken a test including the following multiple-choice item.

The United States Constitution went into effect in (1) 1774,
(2) 1776, (3) 1783, (4) 1789, (5) 1793.

Of these 443 pupils, 303 or 65 per cent selected the correct response to item 4 in the matching exercise, but of these 303 pupils only 176 had been able to select the correct response to the multiple-choice exercise dealing with the same

item of information. It is clear that a large proportion of the pupils responding correctly to item 4 of the matching exercise did not have the historical information called for but had probably noticed the grammatical consistency existing between "it" and "Constitution," and were thus able to respond correctly.

Illustration 2⁵ (Matching exercise in biology)

- | | |
|--|---|
| 1. Most normally green-plants lose their color when | a. through their stomata |
| 2. The common characteristic of flowering plants is | b. contracts into a rounded mass |
| 3. Almost all plants which form coal | c. grown in the dark |
| 4. When an expanded amoeba is strongly stimulated it | d. are now extinct |
| | e. the formation of a reproductive body |

In this item, the pupils were directed to write before each incomplete statement in the left-hand column the number of the phrase which would make of it a complete and true statement. The exercise is so constructed, however, that for each incomplete statement there is only one phrase which would make of it a grammatically correct sentence. Pupils aware of this fact could make a perfect score with no knowledge of biology whatsoever. What this exercise may really measure, then, is the student's sense of grammatical consistency or of correct sentence structure, rather than his achievement in biology.

⁵ Adapted from illustration on p. 380 of *Traditional Examinations and New Type Tests*, by C. W. Odell (Century Co.).

GENERAL CONSIDERATIONS

2. Clues provided by pat verbal associations or by identical elements

Illustration 3⁶

Railroad companies and other companies carrying on interstate commerce are now regulated by a commission appointed by the president of the United States. This commission is called

- the Civil Service Commission
- the National Chamber of Commerce
- the Interstate Commerce Commission

The less the student thinks about this item, the more likely he is to respond correctly. The words "interstate commerce" and "commission" appearing in both the question and the response are a clue to which even the dullest pupil can hardly fail to respond, even though he may have no appreciation of the meaning of the words.

Illustration 4⁷

Since water or other liquids will flow out of a hole in the side of a containing vessel, liquids must exert

Here the pupil need only note that "exert" is a word usually used with "pressure" in physics, and may thus make the correct response only because he knows these words usually go together.

Illustration 5⁷

The ratio of the velocity of light in air to its velocity in another medium is called the of refraction of the medium.

About all the student needs to know to respond correctly to this item is that there is a concept in physics identified as the "index of refraction." The words "of refraction" follow-

⁶ From *Gregory Test in American History*, Test III, Form B, Part 6, Item 2. C. A. Gregory Co., Cincinnati.

⁷ From a teacher-made test.

ing the blank will alone suggest that "index" should precede them, particularly with the help of the clue "ratio" previously presented. Many students might be expected to complete this item correctly who are unable to define the term involved and do not appreciate its meaning.

Illustration 6⁸

- The "Boxer Rebellion" which occurred in 1900 was
- a rebellion of one of the countries in Mexico in protest against the high taxes being imposed upon it.
 - an attempt of the Filipinos to gain their freedom a short time after we had secured the islands from Spain.
 - an uprising of a group of people in China known as "Boxers" who tried to rid their country of all foreigners by murdering them and taking their property.

Here, as in illustration 3, the correct response is "given away" by the identical elements in the question and response, in this case the word "Boxers."

These illustrations may be sufficient to indicate that care must be taken in item construction to insure that the process of word matching is not alone an adequate basis for selecting the correct response, or that items may not be answered correctly simply on the basis of noting which response is externally most consistent with the way in which the question was raised. Defects of this kind are frequently found in even the better standardized tests and contribute in part to the feeling of suspicion with which the objective type of test is viewed by many examiners. The preceding illustrations, furthermore, do not by any means exhaust the types of irrelevant clues which may be found in objective tests. Other illustrations will be presented in the more specific discussion of separate testing techniques.

⁸ From *Gregory Test in American History*, Test III, Form A, Part 7, Item 5.

3. *Specific determiners.*

Ruch, Weidemann, Brinkemeier, and others have made an intensive study of those characteristics (called "specific determiners" by Weidemann) of true-false statements which will prejudice the response by the pupil apart from the thought content of the item. Among other things, they have found that, among a very large number of true-false statements actually constructed and used by classroom teachers,

Four out of five statements containing "all" were false.

Four out of five statements containing "none" were true.

Nine out of ten statements containing "only" were false.

Three out of four statements containing "generally" were true.

Four out of five "enumeration" statements were true.

Two out of three "reason" or "because" statements were false.

Three out of four statements containing "always" were false.

The longer the statement, the more likely it is to be true.

Obviously, the brighter students are apt to note these tendencies in true-false test construction, and may capitalize on their knowledge in responding to statements containing these determiners in cases where they are unable to respond on the basis of the thought content. For example, if the pupil always marks as *true* the statements containing "none," "generally," and "may," and as *false* those containing "all," "always," "only," and "because," he is likely to make a much higher score than his true knowledge of the subject tested deserves, unless, of course, these tendencies have been controlled by the constructor of the test. In such cases, a test item or a complete test may be as much a measure of general intelligence, or "brightness," or mental alertness, as of true achievement in the field involved.

Non-Functioning Elements in Test Items

Defects in test items due to the presence of totally irrelevant clues are usually relatively easy to detect and are therefore readily eliminated by anyone who is at all conscious of the problem. More serious is the type of defect in which the specific element which determines the pupil's response is a part of, rather than totally irrelevant to, the intended content. Such defects are often quite subtle in character, and their presence in even some of the most widely used standardized tests appears to have been overlooked.

Illustration 7⁹

Directions: Below are two columns of items. Match the items in the two columns by placing on the line before each group of words in Column A the right *number* from Column B.

Column A	Column B
..... 1. a Phoenician contribution to civilization.	1. Mason and Dixon Line
..... 2. most famous building of the ancient Greek world.	2. Spanish Armada
..... 3. the fleet whose defeat in 1588 gave England the control of the Atlantic Ocean.	3. Saratoga
..... 4. a boundary between two colonies that later became famous as the division between free and slave territory.	4. Dred Scott Decision
..... 5. the victory which caused France to come to our aid during the Revolutionary War.	5. Parthenon
	6. Missouri Compromise
	7. Alphabet
	8. Printing press
	9. Ordinance of 1787

⁹ From course of study test in American history, Kansas City, Mo., Public Schools, May, 1931, Part II, Group V.

GENERAL CONSIDERATIONS

- 6. the law that forbade slavery north of the Ohio River.
- 7. a ruling by the Supreme Court which opened all territory to slavery.

This is an exercise in which only the most superficial type of knowledge of history plus a certain degree of mental alertness is required if the pupil is to make a high score. In responding to item 4, for example, the student needs only to recognize that only one term in column B could conceivably be the name of a boundary, and the choice is made particularly easy by the fact that the correct response contains "line," which directly suggests "boundary." For item 1 there are only two plausible responses, and since "alphabet" obviously preceded "printing press," the former is therefore likely to be chosen by the pupil who knows only that the Phoenician is an ancient civilization and who can recognize that no other responses in column B look like "contributions to civilization." In item 7, "ruling" and "supreme court" suggest "decision" in response 4. In each case, then, the student has only to recognize what is the *general* nature of the response called for, whether it is a law or a ruling or a name of a building, etc. Since in no case are there more than two responses with the required general characteristics, and in some cases only one, his choice is made relatively easy if he is intelligent enough to recognize these clues. Certain and specific information on his part is not required.

This matching exercise, again, was administered to a group of 433 pupils who had previously taken a test including a number of multiple-choice exercises, each of which tested independently the information called for in one item of the matching exercise. In these multiple-choice items the possibility of responding correctly on the superficial bases just

THE THEORY OF TEST CONSTRUCTION

noted was eliminated or considerably reduced by using wrong responses which were homogeneous in nature with the correct response. For example, the multiple-choice item paired with item 7 in the matching exercise was as follows:

All of the territories were opened to slavery by the (1) Wilmot Proviso, (2) Compromise of 1850, (3) Kansas-Nebraska Act, (4) Dred Scott Decision, (5) Missouri Compromise.

The following table shows the number of pupils who responded correctly to each matching item and, of those pupils, the number who also responded correctly to the multiple-choice items. It may be noted from this table, for example, that less than half of the pupils who responded correctly to item 7 in the matching exercise had really acquired the information called for, and that the majority of them were responding on a superficial basis, presumably having noticed that only one of the items in column B suggested a ruling of the Supreme Court.

NUMBER OF CORRECT RESPONSES

	Matching	Multiple-Choice
Item 1	413	345
Item 2	440	255
Item 3	437	389
Item 4	421	96
Item 5	436	269
Item 6	190	111
Item 7	280	107

Illustration 8

The Frenchman who first explored Lake Michigan was
(1) Kosciusko, (2) Nicolet, (3) Raleigh, (4) Gilbert,
(5) San Martin

This item was administered to 444 pupils who had previously responded to the following item:

He first explored Lake Michigan. (1) Rochambeau, (2) Car-tier, (3) Duquesne, (4) Genet, (5) Nicolet, (6) Champlain.

GENERAL CONSIDERATIONS

Of the 259 pupils who responded correctly to the first of these items, only 66 responded correctly to the second. Clearly, the first item was made easy to any student who was able to recognize that four of the responses were obviously not French names but who would otherwise have been unable to recall or to recognize the correct response.

Illustration 9¹⁰

✓ At the opening of the Washington Disarmament Conference, Secretary of State Hughes proposed a "naval holiday" of ten years in the construction of battleships. At this conference also, Japan agreed to withdraw from Shantung. In what year was this conference held? (1) 1876, (2) 1890, (3) 1914, (4) 1921, (5) 1928. ✓

✓ This item may appear to hold the student responsible for a considerable amount of important information, but he cannot fail to respond correctly if he knows only that Hughes was Secretary of State in 1921 and not in any of the other years given.

Of 174 pupils, out of 444, who responded correctly to this item, only 107 and 103 respectively were able to respond correctly to the following items:

✓ The Washington Naval Conference began its sittings in (1) 1890, (2) 1905, (3) 1914, (4) 1921, (5) 1928.

The question of a "naval holiday" for several world powers ✓ was discussed by their delegates in (1) 1890, (2) 1905, (3) 1914, (4) 1921, (5) 1928.

Illustration 10

The leader in the making of the compromise tariff of 1833 was (1) Clay, (2) Webster, (3) Jackson, (4) Taylor, (5) Harrison.

¹⁰ Adapted from *Iowa Every-Pupil Test in American History*, 1931.

THE THEORY OF TEST CONSTRUCTION

Of 443 pupils responding to this item, 278 selected the correct answer. It would appear, however, that the majority of these students responded on the superficial basis of a strong verbal association between the words "Clay" and "compromise." When these 278 students were given the following item, in which the "compromise" clue was eliminated, only 132 responded correctly.

✓ The leader in the tariff revision of 1833 was (1) Clay, (2) Webster, (3) Jackson, (4) Taylor, (5) Harrison.

Of the 278 who responded correctly to the first item, however, 248 were apparently in possession of the "compromise" clue, since that number responded correctly to the following item:

Who was known as the great compromise-maker? (1) Clay, (2) Webster, (3) Jackson, (4) Taylor, (5) Harrison.

Illustration II¹²

✓ The Monroe Doctrine is

- a law, passed by Congress during Monroe's administration, stating in substance that the American continents are not open for future colonization by European nations and that any attempt at colonization or at re-subjecting nations now free would be considered an unfriendly act.
- not a law but simply a declaration of our foreign policy made by Monroe in his message to Congress.
- a theoretical form of government proposed by Monroe but rejected by Congress because of its being unconstitutional.

Again an item that has excellent "content," but very little of which necessarily functions. If the student reads only the first words in each response he can select the right answer if he knows only that the Monroe Doctrine was neither a law nor a form of government. ✓ With this superficial information

¹² From *Gregory Test in American History*, Test III, Form A, page 5, Item 3.

GENERAL CONSIDERATIONS

he could respond correctly without reading at all any of the latter part of each response.

Illustration 12¹²

- | | |
|--|----------------------|
| () 66. Determined the speed of light recently. | (1) Index of |
| () 67. The property of light that varies inversely. | refraction |
| () 68. The intensity of illumination of a candlepower light at a distance of one foot. | (2) The foot candle |
| () 69. The ratio of the velocity of light in the air and its velocity in any medium. | (3) Michelson |
| () 70. What happens when light tends to pass from a denser to a rarer medium at an angle greater than the critical angle. | (4) Total reflection |
| | (5) Intensity |
| | (6) Faraday |

In this exercise, item 66 obviously calls for the name of a man. Only two proper names are given. The pupil who knows only that Michelson is a present-day scientist ("recently") and that Faraday is not can easily make the correct selection. Item 67 calls for a "property"; "intensity" is the only property listed. In item 68 the words "candle" and "foot" strongly suggest response 2, "the foot candle." In item 69 "ratio" suggests "index." For item 70, "total reflection" is the only response that "seems to fit" the form of the question: "what happens when. . ."

Illustration 13¹³

Factor the following to determine its roots; then underline the correct response.

$$3x^2 - 4x - 4 = 0$$

- (a) 5, 4 (b) -2, -2 (c) 2, $-\frac{2}{3}$ (d) $\frac{2}{3}$, -2 (e) 3, -1

¹² From physics test of Indiana State High School Testing Program of 1931, Form A, page 4, Items 66-70.

¹³ From teacher-made test in high-school algebra.

This item was intended to test the student's ability to factor a quadratic equation. It is quite probable, however, that many students would select the correct response by *substituting* the various values given to find the pair that satisfies the equation. In that case the item would measure the student's ability in substitution plus his ability to recognize the possibility of this type of solution, rather than his factoring ability. Recognition tests in mathematics frequently present opportunities for the student to work backwards from the answers given to select the correct response by a *checking* process, rather than by the process which it was intended to test.

While further illustrations could be supplied without limit, those presented should be sufficient to identify the general problem. Test exercises of the objective type in general require the student to *recognize* as correct or incorrect a response whose form and phrasing are supplied to him, rather than to provide the form and phrasing of that response himself, on the basis of recall, inferential thinking, or the application of some skill. Because of the nature and prominence of this recognition factor, the objective test exercise is particularly susceptible to weaknesses resulting from the unintentional inclusion of irrelevant clues, or from the possibility that the student may base his selection on a recognition of a part, rather than the whole, of the response provided. The particular or peculiar phrasing of a test exercise, its structure, mechanical or typographical arrangement, and other external features may thus play an unsuspected but significant rôle in determining its validity.

One of the most important abilities necessary for objective test construction, therefore, is the ability to anticipate the specific mental processes of the student in reacting to each individual item. The test author must be able to put him-

GENERAL CONSIDERATIONS

self in the place of the student and to analyze the kind of thinking that the student is likely to do. If an analysis of this type reveals that an item in its original form contains clues which are likely to enable the student to respond correctly on an irrelevant basis, or if there is any probability that the student will react only to a minor element in the item, then the phrasing and arrangement of the item must be revised accordingly. The following questions, therefore, should be uppermost in the teacher's mind in constructing a test exercise:

1. Is there any basis, other than that intended, on which the pupil can respond correctly to this item?
2. What is the minimum amount of information, skill, or understanding that the pupil must possess in order to respond correctly? How does this minimum compare with the intended purpose of the item?
3. Is there any element in this item that the student may disregard entirely and yet respond correctly? Are there any elements that need not necessarily function in determining the student's response?]

In most tests of the objective type which have been constructed by the classroom teacher, these questions probably have never been considered.

It is partly because of such technical weaknesses that so many students prefer the objective type of examination. The brighter students, particularly, have discovered through experience that they can often make high scores on such tests without specific and thorough preparation for them. They of course do not dissect or analyze the items or their own reactions to them as has been done in the preceding illustrations, nor do they frequently set out on a conscious and deliberate hunt for irrelevant clues or specific determiners. What they do is to base their responses on unanalyzed "hunches" or

THE THEORY OF TEST CONSTRUCTION

guesses, but very often clues of the kind that have been pointed out are the basis of these "hunches." Students who thus score high on such defective tests are nearly always the bright or smart pupils but not necessarily the good students. Tests of the objective type which exhibit these defects are often as much a measure of "smartness" or general intelligence as of sound understanding or real achievement.

Finally, it is important to note, in fairness to the objective testing idea, that defects of these types are not inherent or unavoidable in objective testing, and that to condemn the objective test as a type because of the prevalence of these defects in the past would be distinctly unjust. Tests of the objective type can be constructed so as to be entirely free from these weaknesses; the fact that they have not been is the fault of the persons who have built them rather than of the type of test employed. These considerations, therefore, should have little or no bearing upon the question of the relative potential values of one type of test as compared to another.

THE NEW-TYPE TEST AND ROTE LEARNING

Among the most frequently recurring of the charges that have been brought against the new-type test are that it has tended to hold the student responsible only for detailed and isolated facts or items of information, that it has in many cases required only rote learning rather than real understanding of this information, and that it has not measured or required from the student the ability, in any significant degree, to organize and integrate ideas, to see relationships, to draw inferences, or to make applications. These charges have not been quite fair to the objective examination as a type, since again the fault has been primarily with the persons who have built the test rather than with the type of test used.

GENERAL CONSIDERATIONS

Furthermore, these charges may often in actual practice be as fairly raised against the traditional essay examination as against the objective test. Regardless of where the blame lies, however, and regardless of the relative merits of the two types of examinations, the charges themselves have undoubtedly had ample justification. ✓ The majority of objective examinations which have been built by classroom teachers for informal use, and many standardized tests as well, *have* tended to be unduly, if not almost exclusively, informational in character, and have tended to place a premium upon rote learning rather than upon true understanding. ✓

The reasons for this tendency are not difficult to understand. In the first place, any teacher-made test will tend to reflect the character of the instruction that has preceded it. It is unfortunately true that a large share of present-day instruction at the secondary school level, and in the college as well, is characterized by close dependence upon the textbook, by much "lesson learning" of the memory type, and by mechanical drill and "recitation" procedures. Such instruction has placed a distinct premium upon rote learning of *statements* of facts and of principles; upon the learning of pat answers to pat questions; upon conscientious memorization of the unique phrasing of the textbook or of the lecture "notes"; and upon conformity with the opinions of the textbook author or of the instructor. It is not surprising, therefore, to find that the tests which have been used in connection with such instruction have been almost exclusively "informational" in character, in the sense that they have failed to measure the student's real understanding of the subject or his ability to do inferential thinking in the field to be tested. ✓ Fundamentally, therefore, the criticisms which have been noted are in a larger degree criticisms of teaching than they are of testing practices. ✓ Until the quality of instruction has been improved in these respects,

it is hardly to be hoped that teacher-made tests will be free from these criticisms.

While much of the improvement now demanded in testing must await similar improvements in the content and methods of teaching, it is by no means implied that a significant degree of independent improvement in test construction is not possible under present conditions. While instruction has been unduly informational in character, testing has perhaps exhibited even more objectionable characteristics in this respect. While teaching has relied too much upon the textbook, the dependence on the text in test construction has probably been even more serious in its consequences. While the quality of testing has, in general, tended to reflect the quality of teaching, there have been many individual instances in which the quality of testing has been decidedly inferior to that of instruction. In other words, there are many individual teachers whose instruction has been of good quality but who have not yet succeeded in effecting a corresponding improvement in their testing procedures. It is to such teachers particularly that the subsequent discussion will prove of most value.

We may note, first of all, that most current teacher-made objective examinations are to be criticized, not so much because they have been concerned largely with the acquisition of "information" *per se*, as because the specific information for which they have held the student responsible has been poorly selected and because of the manner of testing for it. The subject matter of any course of instruction, particularly in the content subjects, may be considered as consisting of a body of purely descriptive or encyclopedic facts, of relationships between these facts, and of ideas, generalizations, principles, and methods of procedure based upon these facts and relationships. All of this may, in a sense, be thought of as "information." Statements expressing a generalization of a relationship are as

much statements of fact as those which are purely descriptive. It is as much a fact that "to every action there is an equal and opposite reaction" as it is that "the specific gravity of mercury is 13.6." Both of these facts may be memorized by the student without any guarantee that he has understood either. Both may be equally meaningful or equally meaningless to him.✓

✓Information, then, may be roughly classified into two types: descriptive facts and interpretative ideas, the latter category including any relationships, generalizations, etc., which are based upon the descriptive material. Of these two types of information, there can be little question that the latter is of more significance and value.✓ There is no field of instruction in which the acquisition of descriptive information constitutes an end in itself. ✓Facts are not learned for their own sake, but because it is assumed that they will prove of functional value in assisting the student to interpret, to understand, to appreciate, or to modify his own environment. Facts of the purely encyclopedic or descriptive variety, furthermore, are less likely to prove of functional value than interpretative ideas.✓ In an analysis of the content of geography as a school subject, we find the following statements:¹

It is obvious that *ideas of relationships* between man and his natural environment are those the value of which (in promoting general educational aims) has been evidenced most frequently in the comments analyzed. The insignificant proportion of encyclopedic facts cited (as of value) is as striking as the large proportion of . . . relationship ideas.

In a questionnaire study of the values of geographical offerings, 83 per cent of those reacting asserted that the greatest contribution geography had made to their ability to solve various problems had been made by *ideas of relationships* between man and his natural environment . . .

¹ From the *Thirty-Second Yearbook of the National Society for the Study of Education* (1933), ch. vii, "Investigating the Value of Geographic Offering," by Edith Putnam Parker, pages 85, 89-91. Quoted by permission of the Society.

Isolated descriptive facts and mere memorization of them were condemned on the general grounds (1) that in most cases such facts were not retained, (2) that mere memorization accordingly was wasteful of time and effort, (3) that in gaining interpretative ideas one came into possession of descriptive facts in such a way that they were retained, and (4) that if knowledge of a descriptive fact that one had not acquired thus were needed, it could readily be gained if one knew how to use source materials effectively.

The findings... point clearly to the relatively great value of interpretative ideas... and to the slight value of mere encyclopedic facts... and of the memorization of such facts.

If the various studies and analyses that were made had led to conflicting ideas of the value of geographical ideas in general education, a more extended investigation would have been necessary. Since, however, there was pronounced agreement in the findings reached by every method employed, and since these methods represented considerable variety in approach to the problem, it seemed valid to conclude (1) that the value of geography in the training of children depends upon the type of geography involved, (2) that in so far as the material presented is a mere mass of encyclopedic facts its value is negligible and the ideas and learning experiences involved have little or no claim for recognition in a curriculum designed for general education, but (3) that in so far as the material presented is such as to give the learners ideas of an interpretative type, ability to gain such ideas for themselves, and power to use them effectively in practical situations, geography can make very valuable contributions to child training.

Each of these statements can be readily paraphrased to apply to other fields of subject matter. Thus, "it may be valid to conclude (1) that the value of 'information' in any content subject depends upon the *type* (or significance) of information involved, (2) that in so far as the 'information' presented is a mere mass of encyclopedic or descriptive facts its value is negligible, and the ideas and learning experiences involved have little or no claim to recognition in a curriculum designed

GENERAL CONSIDERATIONS

for general education, but (3) that in so far as the 'information' presented is such as to give the learners ideas of an interpretative type, ability to gain such ideas for themselves, and power to use them effectively in practical situations, that information can make very valuable contributions to the training of students."

There is some danger, of course, in carrying this argument too far. The second of the paraphrased statements in the preceding paragraph is certainly not valid as a broad generalization and does not give due recognition to the importance of information as such. While it may be true that the descriptive informational materials which the pupil acquires under current instruction may, after he has completed the course, be neither as well retained nor as frequently used by him as the interpretative materials, descriptive facts are nevertheless essential. Thinking cannot go on in a vacuum, of course, and empty generalizations are no better than none at all. The statement has often been made that the most competent thinker in any field is the one who has at his ready command the most extensive information and whose information is most exact. This statement might perhaps better be revised to: the best thinker is the one who not only has acquired the greatest amount of exact information, but who best understands the meaning and significance of, and the relationships between, the facts at his command, and who can best use these facts and relationships. Every descriptive fact, in order to be useful, must be accompanied by an understanding or appreciation of its significance and relationships. In other words, it must be accompanied by interpretative ideas. The real point then is, not that descriptive facts should not be taught or tested, but rather that teaching and testing should not hold the student responsible *only* for such facts. Both types of material have their legitimate place in both testing

and teaching; the real problem is that of securing the proper balance between them.

Unfortunately, the items of information for which students have been held responsible in most teacher-made tests of the objective type have been almost exclusively of the descriptive variety. In the social studies, for example, the typical objective examination has consisted almost exclusively of exercises of the *who*, *what*, *when*, and *where* varieties, requiring the student to associate names of events, personages, locations, and institutions with brief descriptive phrases, or requiring verbal associations between men and events, events and locations, dates and events, terms and definitions, etc. Questions of the *how*, *why*, *with what consequences*, or *of what significance* types, requiring the student to provide, recall, or recognize acceptable *interpretations* of events, institutions, and practices, have been in the decided minority or have been entirely lacking. In the physical and natural sciences the great majority of objective test exercises have required the student to recall or recognize correct *definitions* of terms or units, formal *statements* of laws or principles, literal formulas, and descriptive facts and numerical constants. Questions in which the student must recall, recognize, or provide acceptable explanations of natural phenomena, or apply laws and principles in new situations, have constituted a significant proportion of the items in very few tests in these fields. In literature, the typical objective test exercise is either the "naming" type, requiring the student to associate authors with names of literary selections, or the type testing for the acquisition of biographical information, or for the recall of details of plot and action, names of principal characters, or characteristic features of important literary selections.

This undesirable emphasis upon non-interpretative materials is perhaps a natural consequence of the fact that most

GENERAL CONSIDERATIONS

objective testing techniques are so readily adapted to such materials. There has been a strong tendency to limit the exercises in objective tests to items of the "short response" types, i.e., to items calling for the recall or recognition of a response consisting of a single word, number, date, or brief stereotyped phrase. It is frequently very difficult, however, to build items of the interpretative type in this form; the testing of interpretative materials usually requires the use of items in which the correct response consists of a relatively long phrase, a sentence, or even a complete paragraph. The construction of such items requires, in general, much more ingenuity and skill in phrasing, much more thorough understanding of or imaginative insight into the subject matter to be tested and into the student's way of thinking, and much more time, than is required for the construction of short-response non-interpretative items. It is relatively easy, for example, to build an objective test item that will discover if the student has learned *who* invented the cotton gin, or even *when* and *where* it was invented, or *what* it was, but it is relatively difficult to build an item that will discover if the student understands and appreciates the tremendous significance of the cotton gin in the development of the South. Almost anyone can in a short time build an objective exercise requiring the student to "match" the names of Boyle, Archimedes, and Pascal with formal statements of the laws and principles which they formulated, but it requires more than the usual amount of skill in test construction and understanding of physics to build items that require the student to supply or recognize an adequate interpretation of a natural phenomenon involving the application of these laws. Consider, for example, the relative ingenuity in conceiving and phrasing plausible alternates, the relative degrees of insight and imagination, and the relative amounts of time required

THE THEORY OF TEST CONSTRUCTION

to construct the first and last items in each of the following pairs.

- ✓1. Burgoyne surrendered at in the year
2. Burgoyne's defeat was a very significant event in the Revolutionary War because it
 - (1) made the British realize that the war was lost
 - (2) was the last great battle in the Revolutionary War
 - (3) made the British more determined than ever to humble the colonists
 - (4) enabled Washington to capture New York City
 - (5) convinced France that an alliance with the colonies against England would be advantageous ✓
1. The law expressing the relationship between the temperature and volume of an enclosed gas at constant pressure was first formulated by (1) Charles (2) Newton (3) Kepler (4) Galileo (5) Hooke
2. Why does a gas expand or exert a greater pressure when its temperature is increased?
 - (1) The molecules of the gas strike each other with greater force
 - (2) The heat occupies the space between the molecules, and the molecules have to spread out
 - (3) The gas holds more water vapor at higher temperatures
 - (4) The heat expands the molecules and thereby causes expansion or greater pressure

The time required and the difficulties involved in the construction of items of the interpretative type may in part explain, but certainly cannot justify, their exclusion from teacher-made or standardized achievement tests. If understanding of and ability to use interpretative ideas constitute important outcomes of instruction, then tests can yield highly valid measures of achievement only if they consist in large part of items of the interpretative type. The present over-emphasis on non-interpretative materials, furthermore, in-

GENERAL CONSIDERATIONS

volves far more than the matter of valid testing only. Tests of the exclusively non-interpretative type will not only fail to measure the really important outcomes of good teaching, but, much worse, they may actually encourage the wrong kind of learning on the part of the student. It is a matter of common knowledge that students frequently say that they study differently for objective tests than they do for essay examinations. Usually they insist on knowing in advance what kind of test they are going to get, in order that they may do their studying accordingly. One reason for this is that they have recognized through experience that tests of the objective type usually have the characteristics which have been here noted. Knowing in advance that they will be held responsible for isolated descriptive facts, they will do their best to memorize those facts, even though they may neither know why the facts are significant enough to deserve learning, nor appreciate the functional values of the information acquired. If they know that only descriptive facts will be called for—that they will not be held responsible for interpretative ideas—they will naturally tend to neglect interpretations and applications in their studying. Study and review in preparation for objective achievement examinations has consequently tended to be of the “skim the textbook” or “cram a lot of facts” type.

This unfortunate tendency in learning has perhaps been further accentuated by other practices which have frequently characterized informal objective test construction, notably that of relying too much upon the textbook for the phrasing of items or in other ways giving the advantage to the conscientious rote learner in test performance. All too many objective tests have been built by a process which consists essentially of simply skimming the textbook and selecting, more or less at random, statements of fact or opinion which

THE THEORY OF TEST CONSTRUCTION

can be used almost intact as test items, perhaps with minor changes to make half of them false for a true-false test, or with crucial words omitted to form a completion test. Such procedures, or any other procedure that encourages the student to rely upon the unique phrasing employed either by the instructor or in the text, are clearly undesirable in testing.✓

In a study by Scott and Myers of pupil understanding of historical terms, the following conclusion was reached.

By parrot mastery of a considerable number of phrases and sentences we can with very little wisdom speak and write what seems to be wise. One may conclude that children have very vague and incorrect notions of some of the terms frequently used by them in their routine procedure. ✓Apparently the child by routine word mechanics can, without knowing very much, seem to know a great deal; he may make a "perfect" recitation without knowing what he is reciting about. A "correct" answer is no proof that a child knows what he has answered.✓

What was true of the elementary school pupils studied in this investigation may be and often is true, in varying degrees, of high school and college students. Where the learning of pat answers to pat questions has been encouraged, students can often quote glib and letter-perfect answers, supplied by the instructor or the text, in response to questions stated in familiar form, but when exactly the same question is put to them in another form they may fail to respond correctly or at all, or when essentially the same answer is differently phrased, they may fail to recognize it as correct.

Consider the following illustrations. The items given below were included in a battery of tests administered to a random sample of 325 physics students in Iowa high schools.

1. What is the heat of fusion of ice in calories?
(Answered correctly by 75 per cent of the pupils.)

GENERAL CONSIDERATIONS

2. How much heat is needed to melt one gram of ice at 0°C .?
(Answered correctly by 70 per cent of the pupils.)
3. Write a definition of heat of fusion.
(Answered correctly by 50 per cent of the pupils.)
4. The water in a certain container would give off 800 calories of heat in cooling to 0°C . If 800 grams of ice are placed in the water, the heat from the water will melt
 - (1) All the ice
 - (2) About 10 grams of ice
 - (3) Nearly all the ice
 - (4) Between 1 and 2 grams of ice(Answered correctly by 35 per cent of the pupils.)
5. In which of the following situations has the number of calories exactly equal to the heat of fusion of the substance in question been applied?
 - (1) Ice at 0°C . is changed to water at 10°C .
 - (2) Water at 100°C . is changed to steam at 100°C .
 - (3) Steam at 100°C . is changed to water at 100°C .
 - (4) Frozen alcohol at -130°C . is changed to liquid alcohol at -130°C .(Answered correctly by 34 per cent of the pupils.)

It will be noted that these items progressively call for more and more thorough understanding of the heat of fusion of ice. Item 1 requires only a verbal association between "heat of fusion of ice" and "80 calories." This is the sort of association upon which physics pupils are frequently drilled in a more or less mechanical fashion until the association is firmly established. The success with which it has been established in this particular case is evidenced by the fact that 75 per cent of the pupils tested gave the correct response to this item.

Item 2 is of essentially the same type as item 1, but employs a different phrasing from the pat form in which the

question is usually stated. Even this slight variation in phrasing resulted in a 5 per cent decrease in the number of correct responses.

The ability to supply the correct answer to either item 1 or 2 clearly can be of no functional value unless the pupil has some notion of the meaning of "heat of fusion." The data from item 3, however, indicate that there were many students who could make the verbal association called for in item 1 or 2 who had no adequate understanding of the meaning of this term. (It may be noted that item 3 was scored in a very liberal fashion, and that many responses were accepted as correct that were technically imperfect.)

Any student who really understood the definition provided in response to item 3 should have no difficulty in responding to items 4 and 5. It will be noted, however, that only 35 and 34 per cent, respectively, of the pupils responded correctly to these latter items. It is clearly apparent from these data that items such as 1, 2, and 3 above can provide only an inadequate basis on which to judge the pupils' understanding of the concept taught. Items of the type of 4 and 5 above are definitely superior. It is significant to observe in this connection that out of the 224 pupils who supplied the correct answer to item 1, only 16 per cent succeeded in all of the remaining items; in other words, only one out of every six students who had acquired the verbal association between "heat of fusion of ice" and "80 calories" had acquired even the low level of understanding of these terms called for in items 2, 3, 4, and 5.

The following series of items is similar to the one just presented, and may be similarly interpreted.

1. Write the formula for Ohm's Law, using letters only.
(Answered correctly by 87 per cent of the pupils.)

GENERAL CONSIDERATIONS

2. Which of the following is a correct formula for Ohm's Law?

- (1) $I = ER$ (3) $R = I \text{ over } E$
(2) $I = E \text{ over } R$ (4) $E = R \text{ over } I$

(Answered correctly by 80 per cent of the pupils.)

3. Which of the following is the correct formula for Ohm's Law?

- (1) $I = R \text{ over } E$ (3) $I = R \text{ times } E$
(2) $R = E \text{ over } I$ (4) $E = R \text{ over } I$

(Answered correctly by 60 per cent of the pupils.)

4. Which of the following is a correct formula for Ohm's Law?

- (1) $I = R \text{ over } E$ (3) $I = ER$
(2) $E = IR$ (4) $E = R \text{ over } I$

(Answered correctly by 56 per cent of the pupils.)

5. Is the expression $.24EIT$ equal to the expression $.24I^2RT$?

- (1) Yes, because $E = IR$
(2) Yes, because $IR = EI$
(3) No, because the formulas do not contain the same letters
(4) No, because EI does not equal I^2R

(Answered correctly by 40 per cent of the pupils.)

6. What causes the fuse in a house lighting circuit to burn out if the two terminals in the light socket come into direct contact with each other?

- (1) The fuse is burned out by the sparks that form in the socket
(2) The closed circuit has a very low resistance
(3) Heat produced in the socket makes the fuse burn out
(4) The closed circuit gets hot because it has a very high resistance

(Answered correctly by 36 per cent of the pupils.)

Items 3 and 4 obviously were confusing to a large proportion of students who had memorized Ohm's Law in only one form. It is significant that only 36 per cent of the students, most of whom had acquired the information called for in the preceding

items, were able to interpret such a simple phenomenon as the blowing out of a fuse in a house lighting circuit. In this series of items, of the 290 pupils who succeeded in item 1, only 18 per cent succeeded in all of the remaining items. Certainly the ability to recall Ohm's Law in its familiar form is not acceptable evidence that the pupil can recognize the law in a less familiar form or use it in an interpretative situation. If an entire test consisted of items similar to item 1, the results might appear to indicate excellent achievement, while, as a matter of fact, the understanding of the pupils might actually be superficial and inadequate.

It is only redundant to say that any information which merely has been learned by rote, whether descriptive or interpretative, can be of no functional value to the student in any situation (except perhaps in some school examination!).
✓ If a test is to rank students in the order in which they understand and can use the facts and ideas acquired, its items must be so constructed that rote learning alone will not be sufficient to enable the students to respond correctly. ✓ Good testing, as well as good teaching, should penalize rote learning, rather than place a premium upon it. A good test in this respect is one in which, among other things, the instructor has assiduously avoided the use of textbook language or of the stereotyped and "catch" phrases or pat verbalizations likely to be acquired by the rote learner. It should be the teacher's objective in test construction so to phrase or present the questions and responses that only a genuine understanding of the concepts involved will enable the student to respond correctly. Unique verbal associations learned by rote by the student should be denied any opportunity to function to his advantage. ✓ New approaches, novel applications and illustrations, and unfamiliar phraseology should be employed whenever possible. ✓ Such practices will offer no serious difficulty

to the good student, but they will and should confuse and handicap the student whose achievement is only superficial.

No further attempt will be made at this point to supply specific illustrations of test items which are good and bad from these points of view. This can be more effectively done in the specific discussions which follow. We may note again that it is our purpose here only to identify in general those issues and problems in test construction which are likely to be met in all fields of subject matter, and to reserve to the later chapters the more specific considerations appropriate to each subject-matter field.

✓In general, then, it is extremely important, not only for the sake of more valid testing, but for the improvement of teaching and learning as well, that we get away from the present overemphasis on non-interpretative materials in objective test construction, that we place a correspondingly greater emphasis upon functional values and interpretative ideas, and above all, that we test for both the descriptive and interpretative materials in such a way that the pupils will not be able to respond correctly on the basis of meaningless verbal associations. ✓ With present tendencies in test construction such as they are, few questions can be raised that are more crucial in the evaluation of an objective achievement examination than those concerning the proportion of interpretative material that it contains and concerning the degree to which the items, individually and collectively, test the student's *reasoned understanding of* and *ability to use* that which he has learned.

While the discussion thus far has been concerned primarily with the general achievement test, it should be noted that the considerations in the last two sections, those dealing with "Apparent vs. Functioning Content of Test Items" and "The New-Type Test and Rote Learning," apply with equal

importance in relation to diagnostic and general achievement test construction.

RELATIVE EFFECTIVENESS OF VARIOUS OBJECTIVE TESTING TECHNIQUES

The problem of the relative effectiveness of various objective testing techniques appears to have been of major concern to test technicians during the last decade. In that time a very large number of studies have been reported which have attempted to determine empirically the comparative validities and reliabilities of the various types of the objective examination, such as the true-false, completion, matching, and multiple-choice types of exercises. Practically all of these studies, however, have been inconclusive, if not definitely misleading, and for that reason no attempt will be made to review them here. The reasons for their inconclusiveness, however, are deserving of some consideration and will be reviewed briefly in the following paragraphs.

1. The studies of comparative validities and reliabilities thus far reported have in most cases attempted to determine the relative effectiveness of the various techniques in general rather than in relation to specific fields of subject matter or in relation to any specific objectives within a given field. There is no more point, however, in arguing the general superiority or inferiority of, for instance, the true-false test and the matching exercise than there is in arguing the general superiority of the objective test and the essay examination. Whatever advantages or superiorities any type of test may have are specific advantages in specific situations. The true-false technique, for example, may be relatively effective in testing for the persistence of popular misconceptions in general science and relatively poor in testing for the ability to factor simple

quadratics in algebra. A simple recall type of test may be quite valid in testing for the student's ability to recall numerical constants in physics but of relatively little or no value in testing for his understanding of complicated interpretative concepts in the same field. Most studies have thus far raised questions in a highly generalized form, such as, "Is the true-false test relatively better than the multiple-choice test?" or "Is the three-response multiple-choice to be preferred to the five-response multiple-choice?" etc. Generalized comparisons of this kind are not only of little or no positive value but may even be definitely misleading, since they may result in a general disapproval of types of tests which in specific instances or for restricted purposes might be highly valuable. In order to be of value to the test constructor, comparative studies must determine the relative effectiveness of various techniques for highly specific purposes. The questions which they raise should be of the following type: "In a given field of subject matter, in relation to what specific objectives within that field may the true-false technique be most effectively used?" or "Which technique of testing is most effective for measuring the comprehensiveness of the high-school student's Latin vocabulary?" Very few studies of the latter type have yet been made. Because of their highly generalized nature, the majority of the studies thus far reported contain little or nothing of value to the test technician in the solution of specific problems.

2. The majority of reported studies of the relative effectiveness of various testing techniques have based their conclusions (concerning the relative "effectiveness" of the techniques investigated) primarily or exclusively upon determinations of comparative reliability or self-consistency of the techniques investigated. Objective descriptions or measures of validity are often extremely difficult to secure because of the lack of

any acceptable independent criterion. Reliability coefficients, on the other hand, can usually be easily determined. For this reason, there has been a tendency to give a great deal of prominence to the concept of reliability, not because of its significance, but because it is so easy to measure and describe quantitatively. The reliability or self-consistency of any test or testing technique, however, is of very minor significance in comparison with the validity of that technique in relation to the specific purposes for which it is intended. The reliabilities of a number of tests intended for the same purpose may, in fact, even be negatively related to their validities; i.e., the test with the highest reliability coefficient may be the least valid, while that with the lowest reliability coefficient may be the most valid. The effectiveness of a given test or testing technique *should* be considered, primarily if not exclusively, as a function of its validity. If one test can be shown to be more valid than another for a given purpose, then it is to be preferred for that purpose, regardless of the comparative reliabilities of the two tests.

3. Again, the majority of comparisons between various techniques of testing that have been reported have failed to control certain important factors that of themselves could readily account for differences which have been found. Among the most important of these factors are:

a. Skill or ingenuity in test construction

Investigator A, for example, who is interested in measuring the amount of scientific information acquired by ninth grade high-school pupils in general science, builds a true-false test and a matching test over the same items of information. In this situation he might show that his true-false test is more reliable and valid than his matching exercise. This may only prove, how-

ever, that A is more ingenious in the construction of true-false tests for this specific purpose than he is in the construction of matching exercises, and it may show nothing at all concerning the relative effectiveness with which these techniques may be employed by other test constructors in the same situation. Another investigator might build a matching exercise over the same items of information that is far more valid for the given purpose than either A's true-false test or matching exercises. The validity of any test in relation to a given purpose is far more a function of the skill or ingenuity evidenced in the application of the technique used than it is of the type of exercise employed. Where this factor of skill or ingenuity in test construction is left uncontrolled, a comparison of the relative validity or reliability of two or more types of tests may show the relative degrees to which the investigator has realized or approached the ultimate possibilities of each type of test in the specific situation involved, but it may not indicate at all how the techniques would have compared had their respective possibilities been fully realized in that situation. There are very few studies in which this factor has even been recognized, and practically none in which it has been adequately controlled. Adequate control of this factor, furthermore, is extremely difficult to secure in the experimental situation. It is quite unlikely, therefore, that empirical studies will in the near future contribute very much to a better evaluation of the various types of objective test exercises. Such studies are valuable for determining the relative effectiveness of specific tests in specific situations, but not for the purpose of establishing generalizations concerning the techniques employed in these specific tests.

b. Administration time

Cook has shown that the comparative validity and reliability of two testing techniques, at least when used with certain types of subject matter, may be largely a function of the time in which each test is administered in the experimental comparison. In most comparative studies of validity and reliability of testing techniques, this factor has not been controlled; or the time of administration has been determined on a completely arbitrary basis, with no assurance whatsoever that the time limits chosen are equally advantageous in the two techniques compared. The following quotation from an article by Lindquist and Cook ¹⁴ will clarify this point.

Both the reliability and the validity of a given achievement test of the pencil and paper type are very often largely a function of the time in which the test is administered, i.e., the time allowed the pupils to complete the test. Obvious as this fact may seem, it is, nevertheless, one which has been disregarded in nearly all research into the relative effectiveness of test techniques. So grievous is this oversight, that, in the opinion of the authors, it renders relatively inconclusive the results of nearly all such researches thus far published. In support of this contention, it is the purpose of this article to discuss the influence of administration time upon test validity and reliability, to introduce and define the concept of optimum administration time, and to suggest a technique for determining this optimum time for any given body of test material.

The usual procedure in research intended to determine the relative effectiveness of two techniques for measuring the same ability may be illustrated by the following hypothetical situation, which may be found duplicated in all its essential characteristics in published articles in almost any field of objective test construction. Let us suppose the experimenter

¹⁴ Lindquist and Cook, *loc. cit.*, pp. 163-64.

desires to determine which of two ways of testing spelling ability, for example, the multiple-choice and right-wrong types of recognition tests, is most valid and reliable. For this purpose he constructs one test of each type, using the same set of basic words in each. After a criterion measure has been secured for a given group of pupils, these two forms are then administered to the same group under controlled conditions, but often in quite arbitrarily determined administration times for each form. Suppose, however, that through preliminary experimentation it has been found that 10 minutes are required for 80 per cent of a sample of pupils to complete the right-wrong test, while 15 minutes are required for the same proportion to complete the multiple-choice form. In the controlled experiment then, the two forms are administered at these predetermined times. Let us suppose that, when administered in 10 minutes, the right-wrong form yields a reliability of .72, while the multiple-choice form, when administered in 15 minutes, yields a reliability of .80. The reliability for the first form is then "stepped up" by means of the Spearman-Brown Prophecy Formula, to the same time limit as was used for the longer form. In this case the "stepped up" reliability for the first form would be .83. The critical comparison is then made in terms of reliability coefficients for equal administration times. (A similar procedure could be followed for validity coefficients, using techniques later discussed in this article, but this has rarely if ever been done in past research.)

The latter part of this procedure appears to take the time factor into consideration, but actually it disregards it at the most critical point. There is no conclusive evidence that the *best* time for administering a test is that which is required for completion by 80 per cent (or any other per cent) of the pupils, nor is there any evidence that this proportion (if it is a valid determiner) does not vary for different types of tests. In the last analysis, then, under the type of procedure described, the time limits for each test are arbitrarily determined. The superiority, in either reliability or validity, thus determined for either form might therefore be reversed if other and equally arbitrary time limits were chosen.

It should be clear, in light of this argument, that no valid

THE THEORY OF TEST CONSTRUCTION

comparison can be made between two test forms until an objective method has been devised for controlling this time factor, i.e., of determining, through preliminary experimentation, the time in which each form should be administered.

The foregoing considerations point to the conclusion that in making a selection from a number of test techniques in any specific test situation or in relation to any specific objective of instruction, the test constructor must, at present, depend almost entirely upon logical considerations rather than upon the experimental or empirical evidence that is now available.

QUESTIONS FOR DISCUSSION

1. What are the two major problems confronting the builder of any educational achievement test? What do you consider to be the relative importance of these problems? Which has been least satisfactorily solved in practice? Criticize or defend the statement at the end of the first paragraph of this chapter.
2. Discuss the importance of the written examination in relation to other means of appraisal of the pupil's achievement.
3. Why should little consideration be given to the question of the relative merits of the "essay" and "objective" types of examinations *in general*?
4. What are the two immediate objectives of nearly all school examinations? List a number of specific uses of examinations, and show that each use depends upon the effectiveness with which these objectives are accomplished. Can you think of any use of tests that does not involve one or both of these objectives?
5. Explain why the concept of test validity is so highly *specific* in character. Why is it meaningless to describe a test as *valid*, apart from any statement of the *specific* purpose to which it is to be put? Show, by original illustration, that the same test may possess many different degrees of validity, depending on the purpose involved in its use.

GENERAL CONSIDERATIONS

6. Distinguish between a *diagnostic* test and a *general achievement* test. Is the total score on a good diagnostic test always a good measure of *general achievement*?
7. Why is the meaningfulness of the single score derived from a general achievement test dependent upon the homogeneity of the field tested? Since no field is perfectly homogeneous, how can the use of general achievement tests be justified where diagnostic tests are also available? Why does the homogeneity of the field have relatively little bearing upon the meaningfulness of results of diagnostic testing in that area?
8. Suppose you were to construct a test of general achievement in spelling, and that you define the "field" to be tested as consisting of a specific list of 4000 words usually included in the elementary school spelling curriculum. Suppose your test is of the ordinary list-dictation type, and consists of 50 words selected *at random* from the list of 4000. Suppose, furthermore, that it is known that pupil A can spell exactly 2000 words from the 4000 and that pupil B can spell 2500. If pupil A takes your test, is he likely to make a score of exactly 25 out of a possible 50? Why or why not? What, in addition to the total number of words he knows how to spell in the list of 4000, will determine his score on your test? How might pupil B make a lower score than pupil A, in spite of B's superior ability?
9. In terms of the preceding illustration, why is it important to have a broad and representative *sampling* of content in a general achievement test? Where practical considerations enforce a highly restricted sampling, why is it important that each individual test item *discriminate* as sharply as possible between good and poor pupils? Why should items of zero difficulty be omitted?
10. How do the considerations concerning item difficulty restrict the test author in his choice of items? Why may he *not* always *distribute* his items evenly over the whole field to be tested? Why are these difficulty considerations of less importance in diagnostic testing?
11. Why are the "per cent" method of scoring and the concept of the "passing grade" inconsistent with the purpose of a gen-

THE THEORY OF TEST CONSTRUCTION

eral achievement test? Apart from this inconsistency, why should this concept be discarded?

12. What is meant by the *discriminating power* of a general achievement test item? How may two equally difficult items differ in this respect? What are some of the factors which may lower the discriminating power of an item?
13. Find several instances of items, from actual tests, that you believe will fail to discriminate because of insufficient learning. Because of wrong learning.
14. Find at least 10 items, of any type and from actual tests, that you believe contain irrelevant clues or cues to the correct response. Point out the clue and suggest how it might be eliminated in each case.
15. Find at least 10 instances of test items, of any type, that you believe contain non-functioning elements. Point out these elements in each case, and explain why you believe they could be eliminated without influencing the distribution of pupil responses to the items.
16. Criticize or defend the statement that the charge of failure to measure real understanding may be as fairly raised against the traditional essay examination as against the objective test.
17. Select a published standardized test in the content subjects — e.g., in the social studies or the physical sciences — and classify the items according to whether they test for descriptive facts or interpretative ideas. Check in the test also each item, whether descriptive or interpretative in character, to which you believe the pupil could respond correctly on the basis of memorization of stereotyped textbook statements which he understands only very superficially if at all. Criticize the test as a whole from these points of view.
18. What do you consider to be the principal reason that both objective and essay examinations have in the past been so heavily weighted with non-interpretative materials?
19. Find several instances, from actual tests, of items that you believe really test for reasoned understanding or ability to make use of some concept acquired, and explain why you believe

GENERAL CONSIDERATIONS

that pupils are not likely to respond successfully to these items simply on the basis of rote learning.

20. Why are the results of published statistical analyses of tests or test types of relatively little value to the test constructor in deciding which type of test exercise to employ in a specific situation? What are the limitations of most statistical studies of the relative merits of different types of tests?

In this chapter illustrations numbers 1, 3, 6, 7, 8, 9, and 10, on pages 67-77, and the data concerning pupil responses to these items, were taken from E. C. Denny, "An Investigation of the Defects and Weaknesses in Certain Objective Test Items in American History." Unpublished doctoral thesis, State University of Iowa, 1932.

The illustrations on pages 92-94 were taken from Alvin W. Schindler, "The Extent of Rote Learning in Certain Units of High School Physics." Unpublished doctoral thesis, State University of Iowa, 1934.

CHAPTER III

THE CONSTRUCTION OF TESTS

CHAPTER II has been concerned only with the identification, presentation, and discussion of the general significance of certain of the broader issues and major problems which must be considered in the construction of school examinations. In this chapter an attempt will be made to summarize these discussions, to point out their more specific implications for tests of various types, and to consider briefly certain minor issues and problems which have as yet received no direct attention. In order that it may be made as concretely meaningful and readily usable as possible, this summary will be presented in the form of an organized list of specific rules and suggestions for the actual construction of tests of the objective type.

The rules and suggestions first presented are those which are generally applicable in the construction of objective tests of all types, while those in the latter part of the summary are concerned specifically with each of the more familiar and frequently used forms; namely, the simple recall or short-answer test the sentence or paragraph completion test, the multiple-choice test, the matching exercise, and the true-false statement or alternate-response type. There are, of course, many additional forms of the objective test. It would perhaps be possible to identify at least 40 or 50 distinct variations in type among the test forms now in use, including, for example, the rearrangement, proof-reading or correction-of-errors, analogies, pied outline, and multiple-response forms. It does not seem worth while, however, to give specific and detailed consideration to each of these variations separately, since most of them are essentially only slight modifications

or simple combinations of the five types that are to be considered. The reader should have little difficulty in recognizing the transferability of the suggestions offered in relation to these five types to any of the many modifications or combinations of them which he may employ in test construction.

It should be noted that, unless otherwise specified, the suggestions contained in this section will apply in the construction of examinations of both the diagnostic and general achievement types.

General Rules and Suggestions

Selection of the Sampling

1. Draw up an outline or table of specifications, indicating the relative emphasis that will be given to each of the various objectives of instruction. ✓This outline or table of specifications should in most cases be multiple in nature; i.e., it should provide for several independent classifications of the content to be tested, each with reference to a different point of view. ✓In building an American history test, for example, the elements of content might be classified chronologically, or topically, or according to the type of history involved, such as social, economic, political, and cultural history, or according to types of associations, such as between men and events, events and locations, dates and events, events and consequences, events and interpretative ideas, men and characteristics, men and accomplishments, historical terms and meanings, etc. Clearly, the items in a test might be satisfactorily distributed with reference to one of these classifications, but unsatisfactorily with reference to another. They might, for example, show the proper distribution in relation to the various chrono-

logical periods studied, and yet overemphasize military and political as opposed to economic and cultural history. There may thus be several independent bases with reference to which the subject matter or the test items may be classified. The relative emphasis, or number of items, to be given to each category should be roughly predetermined for each classification. Do not make the usual mistake of relying upon a topical outline only!

2. As each item is prepared in tentative form, tally it in the appropriate category in each classification. (The same item, of course, may sometimes be tallied in more than one category in the same classification.) Attempt to maintain the proper distribution of items with reference to each classification according to the predetermined proportions.

Construction of the Individual Items

1. Do not attempt to determine in advance what type or types of items to employ or what proportion of items will be of each type. In other words, do not start out, for example, to build a test one-half of which is to consist of true-false statements and the other half multiple-choice, or with any other predetermined plan. In relation to each element to be tested, raise the question, "What is the best technique of testing to employ in this situation?" and then build the item originally in that form. (Do not hesitate to employ essay questions where objective techniques do not appear applicable.) Later, when all items of the same type are collected, if there are too few items of one type to justify a separate section, these may be changed into another type already employed. In general, do not use very many different types in any single examination. The suggestion has

frequently been made that the use of a variety of test forms in the same examination is desirable as a device for arousing the pupil's interest. It is extremely doubtful, however, if any such highly superficial features can have any significant influence upon the student's interest in what he is doing, and certainly there are no experimental data to show that a variety of types in a test will of itself improve the student's attitude toward it.

2. ✓ Make certain that each item actually measures what it is intended to measure. In other words, evaluate the item on the basis of its *functioning* content rather than on the basis of its apparent or intended content.¹ If the two do not appear to agree, make the necessary revisions in the form of the item. Perhaps the most important single ability in item construction is the ability to anticipate the specific mental processes of the student in reacting to the item. In making an analysis of this type, raise the following questions:

- a. Does the item contain any irrelevant cues or clues to the correct response?
- b. Does *all* of the item function? What is the minimum amount of information or understanding that will enable the student to respond correctly? Could any part of the item (modifying phrases, qualifications, etc.) be eliminated without significantly influencing the distribution of pupil responses to the item?
- c. Is the *point* of the item sufficiently clear? That is, is it free from ambiguity?
- d. Does the selection or provision of the correct response require from the pupil a real or reasoned *understanding* of the concept tested or only the recall or

¹ See pp. 66-81.

recognition of a unique set of words which may have been memorized without having become meaningful? The use of unique textbook language or of any "catch" phrases or stereotyped verbalizations that may have been employed frequently during the course of instruction is particularly likely to place a premium upon "lesson learning" of the memory type. Tests consisting of pat questions to which equally pat answers have been memorized by the student are not likely to provide valid measures of his understanding or ability to make use of the information memorized. It is therefore best to use new phrasing, new settings, novel illustrations, etc., whenever possible, in an attempt to force the student to a consideration of underlying meanings rather than of the external or verbal characteristics of the responses. A deliberate attempt should be made to penalize rote learning, and an effective device for doing so is to give the *incorrect* responses the external characteristics of the pat answers or stereotyped responses which the pupil may have memorized without understanding them.

3. Avoid the usual overemphasis upon testing for the acquisition of isolated descriptive facts as opposed to testing for understanding of interpretative ideas. ✓ Emphasize the *why*, *wherefore*, *how*, *with what results*, *of what significance*, and the *explain* or *interpret* types of questions in preference to the *who*, *what*, *when*, *where*, and *define*, *describe*, and *name* types of questions. ✓ Test as directly as possible for the *functional* values in what has been learned, and minimize as much as possible the *formal* aspects of instruction. In English, for example, test directly for the student's ability to avoid or to

recognize and correct errors in actual writing situations, rather than only for the ability to recall formal rules, to name parts of speech, etc. The learning of rules and forms is always a means to an end; test directly for the desired end result, without placing an undue premium upon the means by which that end was attained.

4. Make sure that each item is *determinate*, that authorities would agree upon what constitutes the correct response to the item.

Assembly of Items into a Tentative Draft of the Complete Test

If separate measures are to be secured for separate objectives, build the test in sections accordingly, collecting all items of the same type to constitute a separate part within each section. If the test is intended to yield a single-score measure of general achievement, build the test in parts, collecting all items of the same type into a separate part. If there are too few items of any one type to justify a separate part in the test, recast each of these items into the appropriate type among those already employed. (The situations in the separate fields of subject matter in which there is clear justification for the separate measurement of various objectives of instruction will be discussed in the later chapters dealing with individual fields of subject matter.)

Editing the Tentative Draft

1. Recheck carefully the distribution of the items according to the tables of specifications already prepared, and eliminate or substitute items accordingly, when required.
2. In the construction of tests of the general achievement type, make certain that the difficulty of the items, individually and collectively, is adapted to the *specific*

group to be tested. Eliminate or recast any items likely to be missed by all pupils or answered correctly by all pupils in that group. Try to avoid any significant *piling up* of items at any one level of difficulty. Provide for a *range* of difficulty from about 5-20 to 80-95 per cent. ✓ Try to adjust the difficulty of the whole test so that the average score will be about half the possible score, and so that the range of scores for the group tested will extend from near zero to near perfect score. ✓ If the test is properly adjusted in difficulty, however, there should be no zero scores and no perfect scores. There appears to be a current misconception, with many teachers, that if an item is to be "good" it must be difficult. There is, however, no essential relationship between the difficulty of a test item and its functional value in the test (see pages 46-49).

It is extremely hard to estimate the difficulty of an item subjectively in advance of its administration. The distribution of item difficulty in an achievement examination can be adequately controlled only through preliminary try-out of the examination. Such procedures, however, are beyond the facilities of the classroom teacher for informal testing. The teacher will be forced, therefore, to rely upon his subjective estimate of the difficulty of individual items in attempting to observe the preceding suggestions. Because of the fallibility of his judgment, it will perhaps be adequate to classify the items roughly into four categories, such as difficult, above average in difficulty, below average in difficulty, and easy, and to try to secure approximately a uniform distribution of items into these four categories, with the preponderance, if any, in the intermediate categories. Within each section of the test, then, the easy items should

be presented first, then those intermediate in difficulty, and finally the most difficult items. Again, because of the fallibility of subjective opinion, it is perhaps inexpedient to attempt to *rank* the items in their exact order of difficulty.

In the construction of tests of the diagnostic type, these difficulty considerations are of minor significance. The desirability of any item for inclusion in such a test is not necessarily dependent at all upon its difficulty.

3. It is preferable that the items be independent of one another, in the sense that each item would call forth the same response from the student even though administered independently or in a different context. The content of one item should not provide a clue to the correct response to another, nor should any item be of such a character that it cannot be properly interpreted without reference to other items.

Preparation of Final Copy

1. Provide for adequate identification of pupil on title page.
2. Make sure that adequate *directions* are provided to the pupil for the entire test and for each section of it. The following example is illustrative of the type of suggestions which the student should receive, either orally or in printed form on the title page of the test.

General Directions: Do not turn this page until the examiner tells you to do so. This examination consists of three parts and requires 90 minutes of working time. The directions for each part are printed at the beginning of the part. Read them carefully and proceed at once to answer the questions. *Do not spend too much time on any one item; answer the easier questions first*, then return to the harder ones if you have time. There is a time limit for each part. You are not expected to answer all the questions in any part in the time limit, but if you should,

go on to the next part. If you have not finished Part I when the time is up, stop work on that part and proceed at once to Part II. No questions may be asked after the examination has begun.

By exercising careful judgment and making shrewd guesses you may profitably answer questions about which you are not absolutely sure; but since your score will be the number of correct answers diminished by a number proportional to the number of wrong answers, you should avoid answering questions about which you are totally ignorant. Shrewd guessing based on intelligent inference will improve your score, but wild guessing on questions that are entirely unknown to you will waste time which you could better put on other questions in the test, and may result in a large subtraction from the number of your correct answers.²

(Once a given group of pupils have become thoroughly familiar with a given type of test, it is of course no longer necessary to repeat a comprehensive set of directions each time they take that type of test.)

3. Make the mechanical procedure as simple as possible, and provide for sample and practice exercises where needed. Guard against splitting of items from page to page.
4. Pay careful attention to typographical features and to the readability of the test copy to be placed in the student's hands.
5. Make certain that the provisions for the pupil's responses are as adequate and convenient as they can be made, both from the point of view of the pupil in taking the test and from the point of view of scoring the examination. Arrange the provisions for responses in vertical columns wherever possible, for convenience in scoring. Eliminate any occasion for unnecessary writ-

² From *Cooperative American History Test*, Form 1933, published by The Cooperative Test Service, New York City.

GENERAL CONSIDERATIONS

ing on the part of the pupil. Use a code system for indicating the correct response wherever possible. Make sure that the mechanical arrangement or typography of the materials contributes to maximum readability.

6. Decide upon the time allowance. If materials are classified according to objectives, or if comparisons are to be made of part scores, it may be best to time each part separately. Try to adjust the time allowance, except in a rate or speed test, so that at least 75 per cent of the pupils will have time at least to *consider* all items in each section. In general, diagnostic tests should be given more liberal time allowance than general achievement tests.

Preparation of the Scoring Key

Where very large numbers of test papers are to be scored by a small corps of clerks, as in wide-scale testing programs where papers from a number of school buildings or systems are to be scored by a central agency, the preparation of the scoring key and the detailed organization of the scoring procedure deserve very careful consideration. In informal classroom testing, however, where each teacher has to score only the papers from his own classes, a satisfactory key can ordinarily be prepared by simply filling in the correct responses on one of the blank test forms. In scoring, then, the teacher simply lays this key test and the pupil's paper side by side and compares the pupil's responses with those on the key test. This process is of course made more convenient if the responses are arranged on the test paper in vertical columns along the margin of the page, so that the columns of responses on the key can be laid alongside the columns of pupil responses. If the responses are so arranged, it may sometimes prove worth while, if enough papers are to be scored, to cut from each page of the key test the

vertical strip containing the responses, and to mount these strips on cardboard. More elaborate key forms have been prepared, such as "fan" and "stencil" keys, for use with standardized tests and in wide-scale testing. It does not seem worth while to discuss these more elaborate devices here, since, where only a few papers are to be scored, the time required to prepare such keys may often be greater than the time saved in using them. Those interested in these more elaborate procedures may readily secure information and suggestions by examining the keys and scoring directions provided with current standardized tests.

Scoring the Test

The simplest and in nearly all cases the most satisfactory procedure for scoring an objective test is to give a credit of one for each correct response, the total score then being the total number of correct responses. Teachers are often inclined to give extra credit for items which they consider particularly difficult or important. There is no convincing evidence that this practice of weighting certain responses results in enough increase in validity of the total score to justify the added trouble that it involves. Weighted and unweighted scores for the same set of test papers nearly always show nearly perfect correlations, which means that each pupil will receive approximately the same *relative* score regardless of the system of weighting employed. Unless very clear evidence exists in its favor, therefore, the practice of "weighting" responses in proportion to their "importance" may as well be eliminated.

For certain types of recognition tests, it may sometimes prove desirable to correct the scores for "chance" or "guessing." Again, however, the available experimental evidence indicates that scores so corrected are only very slightly, often negligibly, superior in validity and reliability to the uncor-

rected scores. It is the writer's opinion that corrections for guessing are in most instances not worth while in the informal classroom testing situation, except in the case of the true-false or two-response types of test. In wide-scale testing programs, however, where comparisons of scores with general norms are to be made, corrections for guessing may be more readily justified. Instances have been found, in such programs, where teachers have deliberately instructed their pupils to guess at all items that they cannot or do not have time to answer on a rational basis, in the hope that their average score will thereby be improved in relation to other schools that do not resort to this practice. In this situation the correction for guessing should effectively discourage this undesirable practice, and should result in a better attitude of the pupils toward the test.

(More detailed consideration of "correction for guessing" procedures will be given in the later sections dealing with the various types of tests.)

Assigning Letter Grades to Objective Test Scores

The score of a pupil on an objective test ordinarily has *relative* meaning only, i.e., it can usually be evaluated only in relation to the scores made by other pupils on the same test. A score of 40, for example, may represent a relatively high performance on one test and a relatively low performance on another. The meaning of a given score on a given test administered to a given group depends upon the number of and the level and range of difficulty of the test items, and upon the level and range of ability of the group of pupils tested. Scores on different tests may not be directly compared, nor may a score on a single test be independently evaluated on an absolute basis. Particularly, scores on objective tests may not be meaningfully expressed as a per cent of the possible score for

evaluation with reference to the arbitrary per cent standards traditionally employed with essay tests.

It is necessary, therefore, in order conveniently to interpret such objective test scores, to express them in terms of position along a relative scale of standard meaning. The procedure most frequently employed is that of assigning letter grades to the scores, defining each letter grade in relative terms. For example, in the five-letter grading system, the grade of *A* may mean "superior"; *B* "good" or "above average"; *C* "fair," "satisfactory," or "average"; *D* "poor" or "below average"; and *E* (or *F*) "inferior," "very poor" or "failing."

If this method is adopted, the problem then becomes that of determining the score limits corresponding to each letter grade. How high must a score be to receive an *A*? Between what values will scores be given a *B*? Etc.

A variety of procedures have been suggested for transmitting test scores into letter grades. That most commonly employed consists simply of arranging the scores in order of size, and of assigning *A*'s to a given arbitrary per cent at the top of the list, *B*'s to another arbitrary per cent lower along the scale, etc. For example, the upper 7 per cent might be given *A*'s, the next 21 per cent *B*'s, the next 44 per cent *C*'s, the next 21 per cent *D*'s, and the lowest 7 per cent *E*'s or *F*'s. These percentages may of course be arbitrarily adjusted to suit the "standards" of the individual teacher or school.

The most obvious objection to this per cent method is that it automatically results in a fixed proportion of *A*'s, *B*'s, etc., regardless of the form of the distribution of ability (scores) in the group tested. To earn an *A*, a pupil need only score higher than a given per cent of his class; how *far* above the average his score lies does not matter. It is well known, however, that there are some classes that contain less than the usual proportion of able or superior students (in some cases even the

best achievement may not deserve an *A*), and that other classes contain much more than the usual proportion of such students. This proportion varies from class to class and from year to year, and the method of transmitting scores into letter grades should take this fact into consideration. In other words, the proportion of *A*'s assigned should vary from class to class, rather than remain fixed, and similarly with each of the other letter grades.

As has already been pointed out, whether a given score in a given group is to be considered as good, bad, or indifferent depends, in the absence of absolute standards, upon the *level* of ability (average) of the group tested, and upon the *variability* in individual ability about that average. The better methods of transmitting test scores into letter grades are those which recognize this principle, and which determine the score limits of each letter grade in terms of some measure of central tendency, such as the average or median, and some measure of variability, such as the semi-interquartile range, the average deviation, or the standard deviation, of the scores made by the whole group tested. Such methods will result in varying proportions of *A*'s, *B*'s, etc., from group to group, or from class to class, depending upon the form of the distribution of scores for each class.

A number of procedures of this type have been suggested and described in the literature of testing. Only one of these methods will be described here. This method has been selected because of its relative simplicity and ease of application, and because it will accomplish essentially the same results as any of the more complicated procedures, among those that are practicable, that have been suggested. No attempt will be made here to discuss the theoretical considerations underlying the development of this technique. It had best be viewed by the user simply as an empirical procedure, whose principal

justifications are that it works and that it is easy to understand and apply.

The steps involved in the application of this technique are described below. Each step is illustrated concretely by application to the sample problem which is presented on page 123.

1. List the scores in a column in whatever order they come. Arrangement in order of size is not necessary. (See "score" column in sample problem.)
2. Get the sum of the column of scores. (In the sample problem this sum is 3230.)
3. Divide the sum of the scores by the number of scores, carrying the result to only one decimal place. (This result in the sample problem is 115.3.)
4. Round this quotient to the nearest whole number. The result will be the *average* score. (115)
5. Opposite each score, in a second column, write the *difference* between that score and the *average* (deviation from the average). Express all these deviations as positive numbers, i.e., pay no attention to plus and minus signs. (See "deviation" column in problem.)
6. Add these deviations. (696)
7. Divide the sum of the deviations by the number of scores to get the *average deviation*. (Do not carry the result to more than one decimal place either in this or in later operations.) (24.8)
8. Add two times the *average deviation* to the average. The result will be the lower limit of the *A* group. (164.6)
9. Add two-thirds of the *average deviation* to the average. The result will be the lower limit of the *B* group. (131.5)
10. Subtract two-thirds of the *average deviation* from the average. The result will be the lower limit of the *C* group. (98.5)

GENERAL CONSIDERATIONS

11. Subtract two times the *average deviation* from the average.
This result will be the lower limit of the *D* group. (65.4)

The limits of the various letter grade groups are thus defined as follows:

Lower limit of A's = average plus two times the average deviation

Lower limit of B's = average plus two-thirds of the average deviation

Lower limit of C's = average minus two-thirds of the average deviation

Lower limit of D's = average minus two times the average deviation

It will be noted that the distance between limits is the same for each letter-grade interval — one and one third average deviations.

The computational procedures described on p. 121 will work most satisfactorily for relatively small groups. If the number of scores involved is any large number, for example 50 or more, it may be more economical to arrange the scores in a frequency distribution and to compute the average and the average deviation by the so-called "short" method. These procedures may be found described in any standard textbook on educational statistics.

Among the more significant characteristics of the average deviation technique of assigning letter grades are the following:

1. If the form of distribution of scores represents a very close fit to the ideal "normal" curve, the proportion of A's assigned will be between 5 and 6 per cent, of B's between 24 and 25 per cent, of C's between 40 and 41 per cent, of D's between 24 and 25 per cent, and of F's between 5 and 6 per cent. These figures perhaps best indicate the proportions in

THE CONSTRUCTION OF TESTS

Sample Problem

(Based on the scores made on an objective test in English by a class of 28 students)

Scores	Deviations	
120	(120 - 115 =) 5	Average..... = 115.
143	28	Twice average de-
91	(115 - 91 =) 24	viation..... = 49.6 (add)
137	22	Lower limit of A's = 164.6
145	30	
110	5	Average..... = 115.
101	14	Two-thirds average
138	23	deviation..... = 16.5 (add)
33	82	Lower limit of B's = 131.5
135	20	
123	8	Average..... = 115.
87	28	Two-thirds average
173	58	deviation..... = 16.5 (subtract)
110	5	Lower limit of C's = 98.5
87	28	
137	22	Average..... = 115.
59	56	Twice average de-
157	42	viation..... = 49.6 (subtract)
131	16	Lower limit of D's = 65.4
110	5	
89	26	
141	26	
117	2	
145	30	
75	40	
104	11	
96	19	
136	21	
3230 (Sum of Scores)	696 (Sum of deviations)	

$$3230 \div 28 = 115.3 \quad 696 \div 28 = 24.8 \text{ (Average deviation)}$$

Average score = 115

A = 165 and above

B = 132-164 inclusive

C = 99-131 inclusive

D = 66-98 inclusive

F = 65 and below

which the various letter grades will be assigned *in the long run* when this method is employed in actual practice.

2. The proportion of *A*'s, or of any other letter grade, assigned will vary from class to class and from test to test, depending upon the form of the distribution of scores. In distributions containing no outstandingly high scores, no *A*'s at all may be assigned, and in distributions containing no outstandingly low scores, no *F*'s may be assigned. (In the sample problem, for instance, one *A*, ten *B*'s, nine *C*'s, six *D*'s, and two *F*'s were assigned in a class of 28 pupils. This class contained more than the usual proportion of "good" students, relative to the class average, but only one outstandingly high performance was made.)

3. The letter grade which each pupil receives will depend upon *how far* he is above or below the class average, rather than merely upon his rank in class.

4. It follows, as a corollary of the preceding point, that letter grades assigned on this basis are *not* directly comparable from one class to another if the level of ability (average) differs markedly in the two classes. If one class is made up predominantly of unusually able students and the other of poor students, the *same* quality of test performance may be far below average (earn a *D* or *F*) in one class and far above average (earn a *B* or *A*) in the other. An *A* earned in a class of poor pupils, therefore, may represent the same achievement as a *D* or even an *F* earned in a class of very good pupils. Such letter grades, therefore, are directly comparable only for classes at approximately the *same level* of ability or achievement. This limitation characterizes all methods of assigning letter grades that are based upon internal standards.

5. It should be clear, from a consideration of the facts noted above, that the grade of *F* as assigned under this method should be considered simply as indicating a performance

which is "low *relative* to the average performance of the whole class," and *not* necessarily as a "failing" performance. Whether a pupil "passes" or "fails" a given test or course should depend, ideally, upon the relation of his achievement to an arbitrary absolute standard (see p. 36). In every class, of course, there is always a "lowest score," or a performance which is low relative to the class average, but in many of these classes even the poorest pupil may have met the minimum requirements for a "pass" or for promotion. In other classes, the general quality of work may be so poor that even the average achievement may be unsatisfactory, and a half or more of the students may deserve a failing mark. The user of this method, or of any similar method based on internal norms, should therefore be cautioned against a mechanical or over-rigid application of it, and particularly should not unquestioningly accept the lower limit of the D group as the "passing" mark. The position of the passing mark should be set independently in terms of other considerations, and may be considerably more or considerably less than "two average deviations below the class average."

Rules and Suggestions for the Construction of Various Types of Objective Test Exercises

The Simple Recall Exercise

Form:

The literature on testing shows no clear agreement upon exactly what is meant by "simple recall" exercises, particularly as distinguished from "completion" exercises. For the purpose of the present discussion, however, the simple recall exercises will arbitrarily be defined as those appearing in the following forms:

A direct question answerable by a single word, number,

GENERAL CONSIDERATIONS

short phrase, algebraic expression or symbol, with a blank following the question, in which the pupil is to write the correct response. Examples:

Who invented the cotton gin?

How many calories will be required to change 8 grams of ice at 0° C. to steam at 100° C.?

Or a simple declarative statement presented in incomplete form, with a blank at the *end* of the statement, in which the pupil is to write the correct completion. Example:

The cotton gin was invented by

Or a list of terms or statements with which the pupil is to associate other terms or statements, but in which the question for all items is contained in the directions to the exercise. Examples:

Directions: After each of the following inventions write the name of the inventor.

- | | |
|-------------------------|--------------------|
| 1. Cotton gin | 3. Telegraph |
| 2. Sewing machine | 4. Telephone |

Directions: Identify each of the following historical characters very briefly.

- | |
|-------------------------|
| 1. Louis Napoleon |
| 2. Hugo |
| 3. Marshal Ney |
| 4. Robespierre |

Directions: In each space provided write the name of the court having original jurisdiction over the case described.

- | |
|--|
| 1. The French ambassador is sued for a store account |
| 2. A counterfeiter is caught and is being tried |

THE CONSTRUCTION OF TESTS

3. Mr. Jones makes an appeal to obtain a lower tariff rate on the goods he imports.
4. Minnesota vs. Wisconsin
5. An individual brings suit to obtain a claim against the United States
6. A department store in Chicago institutes criminal action against a shoplifter

The following type of question would be based on a map or diagram, on which the various parts or locations had been numbered or lettered.

Directions: Each of the numbers below appears on the accompanying map (or diagram). In the blank opposite each number below write the name of the city, state, river, lake, or other geographical feature it identifies on the map (or write the name — or describe the function of — the part or structure it identifies on the diagram).

- | | |
|---------|---------|
| 1. | 5. |
| 2. | 6. |
| 3. | 7. |
| 4. | 8. |

Possibilities and Limitations:

1. This type of test is well adapted to testing for the acquisition of descriptive information, of verbal associations of the *who*, *what*, *when*, and *where* types, or of the ability to name or to identify things briefly described or characterized. It is particularly valuable in computational problem situations in mathematics and the physical sciences.

2. It is extremely easy to build, and for this reason perhaps

GENERAL CONSIDERATIONS

tends to be excessively used. Its use frequently results in overemphasis upon the acquisition of descriptive facts or upon "naming" and "identification" abilities.

3. Because of the subjectivity involved in scoring the longer responses, its usefulness is greatest with short-response types of questions. It is difficult to use in content subjects in testing for the student's understanding of complicated concepts or for his ability to do inferential thinking; i.e., it is difficult to use in connection with interpretative materials. (Techniques helpful in overcoming these difficulties are suggested in Chapter V.)

Rules and Suggestions:

1. In general, attempt to use this type of item only where the correct response consists of just a single word, number, algebraic expression, symbol, or very brief phrase.

2. Use the direct-question form in preference to the incomplete-statement form wherever possible. Direct questions are less artificial and more familiar to the pupil, can be more readily phrased without ambiguity, are less likely to contain irrelevant clues to the correct response, and can be scored more conveniently.

3. Make minimum use of textbook language or of stereotyped expressions in phrasing the question or the incomplete statement. Minimize the possibility that meaningless verbal associations or sheer word mechanics may enable the student to respond correctly. Use sparingly, if at all, any pat questions to which pat answers may have been learned by rote. Employ novel approaches or unfamiliar phrasing wherever possible.

4. In the incomplete sentence type, make certain that the *kind* of response wanted is clearly indicated. For example:

Columbus was born in

THE CONSTRUCTION OF TESTS

does not indicate whether time or place of birth is desired.
A better phrasing would be

Columbus was born in the year

or

Columbus was born in the city of

5. Avoid the possibility that the grammatical structure of the question or incomplete statement may enable the student to eliminate the wrong responses. For example, avoid calling for a plural response where incorrect responses in the singular might otherwise be provided by the student, and avoid prefacing the blank by a delimiting article such as "a" or "an."

6. Check against the possibility of several correct responses. In general, use only questions for which there is but one correct response or a very limited number of correct responses which may be handled conveniently in a scoring key.

7. Make certain that *all* of the content of the item necessarily *functions*. Can the student respond correctly if he reads only a part of the question or of the incomplete statement, or if he notes only a single word or phrase within the question or statement?

8. The simple recall type of test is often used excessively in testing for the student's ability to *name* something which has been described or defined. In general, the better method of approach in testing for the student's understanding of a term or concept is to start with the name and to require the student to provide the definition or description, or to recognize which of a number of alternate definitions or descriptions is correct. The simple recall type of test is not adapted to the latter type of approach. (See page 145.) Avoid the tendency to use the former type of approach excessively with this type of item.

9. Make adequate and definite provisions for the pupil's

GENERAL CONSIDERATIONS

responses (blanks, parentheses, etc.). Do not force the pupil to decide for himself where on the paper his response is to be written, and allow enough space so that he can write naturally and legibly.

10. In computational problem exercises, as in physics or chemistry, make it clear to the student whether or not he is to indicate the *units* in which the answers are given. In the following illustrations, for example, the (b) and (c) forms are preferable to that of (a). Example:

How much heat is required to raise the temperature of a kilogram of water from 30° C. to 90° C.?

Possible provisions for responses	{	(a)
	calories	(b)
		AnswerUnits	(c)

11. Arrange the provisions for responses (blanks) in a vertical column in the right-hand margin of the page for convenience in scoring.

The Sentence or Paragraph Completion Exercise

Form:

A sentence or paragraph in which certain words or phrases have been deleted and blanks substituted for the student to fill in. Examples:

The of a sound varies as the square of the distance from the source.

Whenever a nation more than it, it is said to have a favorable balance of trade.

In 1795, Napoleon was given command of a small army in The wonderful genius of the young leader made the campaign the decisive factor in the

THE CONSTRUCTION OF TESTS

war. By rapid movements he separated and forces. In eleven days he forced to conclude peace. Then turning upon the brave he won one battle after another. By July he was master of Austria clung to her provinces. During the following year four fresh armies were sent from the Rhine to the and each in turn was defeated. In October, 1797, Austria agreed to accept from Bonaparte in exchange for and, which she had lost. This war closed with the peace of

Modifications: (These are illustrative only; many others could be provided.)

Directions: Complete and balance the following equations.

Example: $\dots \text{H}_2\text{O} + \dots \text{Na} = \dots \text{NaOH} + \dots \text{H}_2$

1. $\dots \text{Zn} + \dots \text{H}_2\text{SO}_4 = \dots + \dots$
2. $\dots \text{CuO} + \dots \text{H}_2 = \dots + \dots$
3. $\dots \text{KOH} + \dots \text{HCl} = \dots + \dots$

Directions: Complete the translation of each of the following sentences.

1. This man is my brother.
(...) homme est mon frère. (...)
2. He is taller than John.
Il est plus grand (...) Jean. (...)

Directions: Write in each blank the correct form of the verb given in the parentheses before the blank.

1. John (run) all the way to school yesterday.
2. Mary has (go) home.

Possibilities and Limitations:

1. The possibilities of the sentence or paragraph completion type of test are very much the same as those of the simple recall type. The claim has often been made that the sentence

GENERAL CONSIDERATIONS

or paragraph completion exercise holds the pupil responsible for the understanding of a *complete* thought and that it encourages integration of ideas. Its usefulness in this respect, however, has been greatly exaggerated. In practice, it tends to be a fact-finding type of test in which verbal associations, knowledge of unique phrasing, and word mechanics play an unduly important rôle.

2. It tends frequently to become a "puzzle" type of item through overmutilation. It is admittedly a good type of item for general intelligence tests, but when used in achievement tests it often tends to measure general intelligence rather than knowledge of a specific subject matter. The pupil may know the correct response but have difficulty in phrasing his idea so that it will fit into the blanks provided.

3. Difficult to score conveniently, because of the "staggered" provisions for the responses.

4. Difficult to score objectively, because of the wide variety of equivalent ways of expressing the same idea.

5. Frequently tends to encourage rote learning of unique and stereotyped statements by the pupil.

6. Perhaps more useful and valuable as a teaching device than as a testing device. May be used to require students to consider facts in relation to one another or in connected discourse.

Rules and Suggestions:

1. Avoid direct copying of sentences or paragraphs from text. This practice places an undesirable premium upon photographic recall and rote learning.

2. Avoid placing an undue premium upon the pupil's ability to supply a *unique* term or phrase where other responses would indicate satisfactory understanding.

3. Avoid *overmutilation*. Omit only *key* words or short

THE CONSTRUCTION OF TESTS

phrases. Make certain that enough of the statement is left to constitute the equivalent of a direct question, i.e., make certain that the *kind* of response desired is clear to the student. Avoid "puzzle" items.

Consider, for example, the following items, which were taken from teacher-made tests. Any reader of this chapter is almost certain to have the information for which these items purport to test. Very few, however, would be able to supply the "correct" completions.

1. Physical education is not necessarily a program, but rather a way of providing profitable to the members of present-day society because it is inherently and to them.
2. Civilized man; uncivilized man

It is significant to note that the majority of the students in the classes to which these tests were given provided the correct completions for the blanks. In the second item, for instance, the pupils recognized that the exercise was based upon the textbook statement, "Civilized man makes things; uncivilized man finds them," which they had memorized. Items of this kind may thus *appear* to work satisfactorily and may therefore escape detection — the pupils are able to respond correctly on the basis of rote learning rather than because they are able to interpret the item properly.

Items of this "puzzle" type are particularly likely to result if the practice is followed of "lifting" sentences intact from the textbook and of eliminating key words. Overmutilation of this kind is not easy to avoid, since the mental set of the test constructor is not conducive to a detection of weaknesses of this kind. The person who is building the item has the complete statement so clearly in mind that it may never occur to him that its specific meaning has been destroyed

GENERAL CONSIDERATIONS

by the omission of the key words. For this reason it is best to have the test criticized by some other person, preferably by someone who is not acquainted with the unique phrasing of the textbook employed.

4. Guard against irrelevant clues to the elimination of incorrect responses on the basis of grammatical consistency or word mechanics. (See illustrations nos. 4 and 5 on page 70.)

5. In general, for the sake of objectivity in scoring, avoid the omission of long phrases.

6. Check against the possibility of many correct responses not anticipated in the scoring key.

7. Do *not* indicate the length of or number of words in the response by the length of or number of blanks provided, or by other similar devices. This practice is only likely to accentuate the weakness with which rule 2 above is concerned. Particularly, do not provide additional blanks for unimportant modifiers such as "a," "an," or "the."

8. Make certain that *all* of the content of the item will actually and necessarily function in determining the pupil's responses.

9. Guard against the possibility that one part of the exercise may suggest or supply the responses called for in other parts. In the following exercise, for example, the pupil need have only a very superficial knowledge of common electrical terms to supply the correct responses for most of the blanks, if only he is sufficiently alert to profit by the internal clues provided. For example, the response to blank 6 is obviously "current," since only a current would be "turned off." The words "opening or closing" before blank 9 readily suggest "switch," apart from the rest of the context of the paragraph. It will be obvious to the reader that most of the rest of the blanks can be similarly

THE CONSTRUCTION OF TESTS

filled by the alert student on the basis of various clues provided, even though he has no specific or certain knowledge concerning electromagnets.

Electromagnets

The experiment with the solenoid teaches us the essential parts of the ____ (1) ____ . These are, a ____ (2) ____ carrying the current and having an iron ____ (3) ____ . Since the magnetism of soft iron is ____ (4) ____ , an ____ (5) ____ with a soft iron core loses most of its magnetism when the ____ (6) ____ is turned off. This makes the electromagnet far more useful than the ____ (7) ____ magnet. They can be magnetized or ____ (8) ____ at will by opening or closing a ____ (9) ____ . Telegraph systems, electric motors, door bells, and many other devices depend upon the action of ____ (10) ____ . We see the tremendous importance of ____ (11) ____ discovery on everyday life. It is important to remember that the ____ (12) ____ of the current, and not the end at which it enters, fixes the ____ (13) ____ of the magnet. Practical magnets have many ____ (14) ____ of wire. The ____ (15) ____ passes through all the wires in the same direction, and they all contribute to the ____ (16) ____ of the magnet.

10. Make certain that the provisions for the student's responses are adequate. Avoid any necessity for crowding that will lead to illegible writing.

11. Greater convenience in scoring may be secured by leaving only short blanks within the connected discourse, numbering each blank, and then directing the student to write each required word opposite the appropriate number in a column at the right of the page, as in the following example.

GENERAL CONSIDERATIONS

In 1795 Napoleon was given command of a small army in 1. The wonderful genius of 1 _____ the young leader made the 2 campaign the 2 _____ decisive factor in the war. By rapid move- 3 _____ ments he separated the 3 and 4 forces. 4 _____ In eleven days he forced 5 to conclude 5 _____ peace. Then turning upon the brave 6, he 6 _____ won one battle after another and by July he 7 _____ was master of 7. Austria clung to her 8 8 _____ provinces.

The disadvantage of this device is that it makes it somewhat more difficult for the student to keep in mind the total meaning of the sentences that he has completed.

Multiple-Choice Exercise

Form:

A direct question followed by a number of responses, only one of which is correct and all others definitely incorrect.
Examples:

1. Who invented the telephone? (1) Morse, (2) Marconi, (3) Henry, (4) Bell, (5) Kelvin.
2. What is the purpose of an electric motor in an electric refrigerator? (1) It compresses a gas and cools it with air currents until the gas is liquefied. (2) It cools the air compartment by blowing air through it. (3) It makes a liquid evaporate by blowing air over it. (4) It produces ice by electrolysis of water.

Or, a direct question followed by a number of responses, all or some of which are acceptable in various degrees but one of which is definitely *better* than any other (best-answer type).

Example:

3. What is the principal advantage of the unicameral over the bicameral form of state legislature? (1) It makes legislative

THE CONSTRUCTION OF TESTS

leadership more responsible. (2) It significantly reduces the amount of money spent for legislative salaries. (3) It enables the legislature to take over extensive administrative duties. (4) It permits less frequent meetings of the legislature.

Or an incomplete statement with several possible completions provided, one of which is to be selected as in the preceding types. Example:

4. The United States entered the World War on the side of the Allies because (1) it was a foregone conclusion that the Allies would win, (2) Germany was the only power interfering with her oceanic shipping, (3) the German blockade unfairly jeopardized the lives of Americans, (4) the Allies promised territorial compensation in the Far East.

Or a list of words or phrases, each of which is followed by a number of words or phrases one of which may be correctly associated with it on some basis indicated. Example:

5. Directions: In each of the following items you are to decide which of the five words or phrases given most nearly corresponds in meaning to the Latin word at the left, and write its number in the parentheses before the Latin word.
- () 1. scio 1 intermitto, 2 succedo, 3 seco, 4 intellego, 5 specto
- () 2. quaero 1 puto, 2 relinquo, 3 rego, 4 reddo, 5 rogo
- () 3. pauci 1 quies, 2 noti, 3 non multi, 4 breves, 5 numerus magnus
- () 4. ago 1 flecto, 2 facio, 3 clamo, 4 factum, 5 augeo.

There are a large number of minor modifications, particularly in typographical form, of the types of exercises illustrated above. The pupil may be directed to underline or to check the correct response, or to copy it in a blank provided. The question may be based on an accompanying map, diagram, or chart, or upon a reading selection or paragraph (in reading comprehension tests in the vernacular or in foreign

languages). The pupil may be directed to select the *least* rather than the most satisfactory response, or to indicate which of a number of terms or phrases do *not* characterize the thing in question. The device has also been employed of including occasional items in which *no* responses are correct, directing the student to indicate this situation by a special symbol, or to write in a blank the correct response. A more significant variation is that in which *more* than one response may be correct, for example "Check *each* of the following contributions to civilization which have been attributed to the Phoenicians." This latter type is characterized as the *multiple-response*, in contrast to the *multiple-choice* type of item. The multiple-response type is difficult to score so as to correct for the increased opportunities for guessing, but it appears to have some very promising applications.

Possibilities and Limitations:

1. The multiple-choice type is perhaps the most valuable and the most generally applicable of all types of test exercises. It can be used in situations in which the simple recall types are inadequate because of the length of or the number of correct responses possible. It can be made particularly effective in requiring inferential reasoning, reasoned understanding, or sound judgment and discrimination on the part of the pupil; it is definitely superior to other types for these purposes.

2. It can be made completely objective in scoring. It can be scored very conveniently, and is well adapted to mechanical scoring procedures.

3. Because of the recognition factors involved, it requires unusual care in test construction to avoid the inclusion of irrelevant clues or the possibility that the student may respond correctly on a very superficial basis.

THE CONSTRUCTION OF TESTS

Rules and Suggestions:

1. The direct question form is most easily phrased and is the most natural form to the pupil. It is less likely to contain ambiguities than the incomplete sentence. Its use usually results in greater homogeneity in the responses. It is less likely to contain irrelevant clues to the correct response. The direct question form, however, tends to be slightly longer in phrasing than the incomplete statement form.

2. When the incomplete sentence structure is employed, have the alternate responses come at the end of, rather than within, the statement.

Poor

Before 1660, (1) indentured servants, (2) Oriental coolies, (3) Negro slaves, (4) Indian slaves, were the chief source of labor supply in the British colonies in America.

Better

Before 1660, the chief source of labor supply in the British colonies of America was (1) indentured servants, (2) Oriental coolies, (3) Negro slaves, (4) Indian slaves.

3. When the incomplete sentence structure is employed, make certain that the incomplete sentence is equivalent to a direct question. All responses should constitute possible answers to a single direct question implied in a complete introductory statement. Avoid a mere collection of unrelated true-false statements. Questions of the latter type have very frequently characterized objective achievement examinations in the past. The item below, for instance, is essentially a collection of true-false statements, all of which happen to begin with the same phrase. There is no single question to which all of the responses in this item could be considered as possible or plausible answers. The second item is definitely superior in this respect.

GENERAL CONSIDERATIONS

1. The Declaration of Independence (1) was drafted by Thomas Jefferson, (2) was signed in 1778, (3) contained an indictment of the English king, (4) was signed by all of the members of the First Continental Congress.
 2. The signing of the Declaration of Independence was significant because (1) it changed the struggle from one of resistance to the unlawful acts of a sovereign to open war, (2) it convinced the Tories of the error of their ways, (3) it represented a reaffirmation of the spirit of the First Continental Congress, (4) it secured the immediate cooperation of France in the Revolutionary struggle.
4. Avoid the inclusion of irrelevant clues to the correct response, involving stereotyped phraseology, placement of correct response, word matching, grammatical consistency, etc.
- a. Avoid the use of textbook language or of pat questions and pat answers. Use an unfamiliar phrasing, both in the question or incomplete sentence and in the correct response, in order to force the student to consider underlying meanings.
 - b. In order to penalize and mislead the rote learner or the shallow thinker, it may be legitimate to use familiar or stereotyped phrasing in an incorrect response, i.e., it may be desirable deliberately to make the incorrect responses look as much as possible like a correct response which may have been learned by rote, or to use as wrong responses stereotyped textbook statements which are correct in themselves but which do not constitute the right answer to the question raised.

In the following item, for example, responses 2 and 3 will be plausible to any student who knows that "evaporation cools" and that moving air hastens evaporation, while response 4 is particularly plausible to the uncritical student be-

THE CONSTRUCTION OF TESTS

cause it contains the words "ice" and "electrolysis," which are so frequently associated with refrigeration and electricity. Among 1000 students tested on this item, 32 per cent selected the correct response (1), 12 per cent response 2, 10 per cent response 3, and 38 per cent response 4.

What is the purpose of an electric motor in an electric refrigerator? (1) It compresses a gas and cools it with air currents until the gas is liquefied. (2) It cools the air compartment by blowing air through it. (3) It makes a liquid evaporate by blowing air over it. (4) It produces ice by electrolysis of water.

In the following item the wrong responses (responses 1, 2, and 4) are all more or less exact textbook statements. Response 2 is particularly plausible because it contains the words "lifting power," which the unthinking student will usually associate with the words "lifting device" in the question. Fifty-two per cent of 1000 students tested chose one of these three wrong responses.

At many service stations, automobiles are lifted to a height of 5 or 6 feet for convenience in greasing. Which of the following principles is applied by the lifting device? (1) A body is buoyed up by a force equal to the weight of the displaced fluid. (2) The lifting power of a lever varies with its mechanical advantage. (3) Pressure applied to an enclosed fluid is transmitted equally in all directions. (4) The gravity pressure of a liquid varies with its depth and density.

c. Avoid making the correct response consistently longer or consistently shorter than the incorrect responses.

d. Avoid the possibility that the student may select the correct response simply because it contains the same word, words, or phrases as the question or as the introductory incomplete statement, or because of other external characteristics. This idea may be used negatively to mislead the rote learner or

GENERAL CONSIDERATIONS

shallow thinker; i.e., words or phrases similar to those in the question itself may be deliberately planted in the wrong responses to give them added plausibility.

e. Make all responses grammatically consistent with the form of the question or incomplete statement.

5. Do not require the pupil to copy or underline the correct response. Use a code system with vertical arrangement of responses for convenience in scoring; i.e., direct the student to write in the blanks or parentheses, arranged in a vertical column, the letter or number corresponding to the correct response.

6. In general, particularly in informal classroom testing, do not use the recognition multiple-choice type of item where the simple recall type is adequate, as follows:

Where there is clearly only *one* correct response and where that response is a single word or number.

Where numerical responses are called for, as in computation problems.

Where obviously there are only two responses that are at all plausible or possible, such as, for example, north pole vs. south pole, positive electricity vs. negative electricity, or clockwise vs. counterclockwise motion in physics.

Where writing the correct response takes no more time than copying the code number corresponding to it.

(Exception: Standardized testing programs, where complete objectivity in scoring is required or where a mechanical scoring system is employed.)

7. In informal classroom testing, there is usually no particular advantage in holding to a fixed number of responses. If only three plausible responses can be contrived, use only three. If five or seven can be provided, provide them. In

general, there is no advantage in using more than seven responses.

(Exception: Where correction for guessing is to be employed. See also exception to preceding rule.)

8. Make certain that the pupil must read all of each response before accepting or rejecting it, i.e., make certain that the entire content of the exercise will necessarily function.

9. Make all responses plausible. Use only responses that have all the external characteristics of the correct response, i.e., make all of the responses homogeneous in their general characteristics and in external form. The object is to make each response so plausible that it will be selected by some students. If a response is not selected by any pupils in the group tested, it obviously has no functional value and might better be eliminated entirely. (Insistence upon a fixed number of responses often results in a large number of such non-functioning responses.)

10. In general, in informal classroom testing it is undesirable to attempt to use any "correction for guessing" formulas. (Exceptions in the case of standardized testing programs have already been noted.)

11. There is some danger, in the construction of recognition items of the multiple-choice type, of making some of the wrong responses so plausible that the item will be negatively discriminating. (See discussion of discriminating power of a test item, pp. 56-65.) In certain cases an incorrect response may appear plausible to the good student because of his positive but insufficient learning or understanding; that is, he may respond incorrectly because he knows something but not enough, while the inferior student may by chance respond correctly as a result of random guessing or because he does not have sufficient knowledge to recognize the plausibility of certain incorrect responses. In general, the test should

GENERAL CONSIDERATIONS

be so constructed that the knowledge or understanding required for the *elimination* of an incorrect response will be on a lower level than that required for the *direct* selection or recognition of the correct response.

12. It should be noted that the multiple-choice exercise tests for the student's ability to eliminate incorrect responses as well as to select the correct response directly. In certain instances it may be well deliberately to make use of this elimination process and to give as much attention to the mental process and understanding required for elimination as to that required for the selection of the correct response.

A type of item that forces the student to consider each response carefully and to arrive at the correct response by elimination and comparison is that in which the student is required to select the least satisfactory response. With this negative approach, it is sometimes possible to make all responses more plausible and to call for more careful consideration from the student than if the best answer form is used, particularly when it is difficult for the test constructor to decide what the best answer would be to a positive form of the question. Consider, for example, the following illustrations:

1. Which is *not* an important result expected from the building of such gigantic dams as those planned for the Colorado (Boulder), the North Platte, and the Upper Mississippi?
 - (1) Flood control
 - (2) Power
 - (3) Irrigation
 - (4) Inland navigation
2. Which of the following principles suggested by President Wilson as necessary for a satisfactory peace settlement following the World War has been *least* realized?
 - (1) Reduction of armaments
 - (2) Formation of a League of Nations
 - (3) Self-determination for subject peoples
 - (4) Establishment of an independent Polish state

13. As has been noted previously in the rules for the construction of simple recall exercises, there are two approaches that may be employed in testing for the student's understanding of a term or concept. With reference to the multiple-choice type of exercise, the first approach is to present a definition or descriptive statement and then to require the student to select the *name* of the thing defined or described from a number of alternate names, while the second is to present the name first, followed by a number of alternate definitions or descriptions from which the student is to make a selection. In general, the latter type of approach is the better and is most likely to defeat the rote learner or the student who relies upon superficial verbal associations.

The first item below, for example, can be readily answered by the pupil on a very superficial basis. He need only know that the definition "looks like" or contains familiar words from a memorized definition of the coefficient of expansion. The second item, however, calls for a thorough and certain knowledge of the meaning of the term tested. In the first type of item the test author is limited, in the selection of wrong responses, to familiar terms in physics. In the second type of item he can make the responses extremely plausible and can call for any degree of discrimination he pleases.

1. The increase in length per unit of length of a metal rod for each degree rise in temperature (Centigrade) is known as
 - (1) the specific heat of the metal
 - (2) the elasticity of the metal
 - (3) the coefficient of linear expansion of the metal
 - (4) the surface tension of the metal.
2. The coefficient of linear expansion of a metal rod is
 - (1) the ratio of its length at 100° C. to its length at 0° C.
 - (2) the increase in the length of the rod when its temperature is raised 1° C.

GENERAL CONSIDERATIONS

- (3) the increase in length when the temperature is raised 1° C. divided by the total length of the rod at the original temperature.
- (4) The rise in temperature (degrees Centigrade) which is necessary to cause a 1 per cent expansion in the length of the rod.

It should be noted that different degrees of understanding may be tested, depending upon the homogeneity of the alternate definitions or descriptions provided, that is, upon the degree of discrimination required to select the correct responses. Consider the following examples (taken from "Constructing Tests in the Social Studies" by Edgar B. Wesley, in University of Iowa Extension Bulletin No. 310, *Aids for History Teachers*):

Engel's law deals with

1. the coinage of money
2. the inevitableness of socialism
3. diminishing returns
4. marginal utility
5. family expenditures.

The pupil is expected to know merely that Engel's law deals with family expenditures. No very specific information is expected of him. The next example affords an illustration of greatly increased specificity.

Engel's law deals with family expenditures for

1. luxuries
2. food
3. clothing
4. rent
5. necessities.

In this instance the pupil is informed that Engel's law deals with family expenditures, and he is asked to select the specific item under this heading.

Let us note a third illustration, which calls for a still higher degree of discrimination:

THE CONSTRUCTION OF TESTS

According to Engel's law, family expenditures for food

1. increase in accordance with the size of the family
2. decrease as income increases
3. require a smaller percentage of an increasing income
4. rise in proportion to income
5. vary with the tastes of families.

In the third illustration one will note that the pupil is informed not only that Engel's law deals with family expenditures, but that it deals with expenditures for food; thus the introduction furnishes more information, but a higher degree of discrimination is required in selecting the right choice. The test-maker must decide in advance as to the degree of specificity, taking into consideration the nature of the item and the amount of training possessed by the pupil.

Matching Exercises

Form:

The matching exercise is fundamentally a group of multiple-choice items, all of which employ the same responses. It may consist of a list of questions or incomplete statements in one column and a list of common alternate responses in another column, or it may consist of a number of descriptive phrases or single terms in one column with a number of terms or phrases in the second column to be associated with those in the first on some basis indicated in the directions to the exercise.

Directions: Write in the parentheses before each statement the number of the location to which it applies.

- | | |
|--|-----------------------------------|
| () 1. The eastern end of this island is located directly south from the most eastern part of the United States. | 1. Bermuda
2. Cuba
3. Haiti |
| () 2. Would be crossed in a direct airplane flight from Key West to the Panama Canal. | 4. Jamaica
5. Nicaragua |
| () 3. The most southern of the places listed. | |

GENERAL CONSIDERATIONS

Directions: In the blank before each work in the left-hand column, write the number, from the right-hand list, corresponding to the name of the author of the work. The same name may be used more than once. Some names may not be used at all.

- | | |
|--|---------------------------|
| 1. Huckleberry Finn | 1. Clemens, Samuel |
| 2. The Deerslayer | 2. Cooper, James Fenimore |
| 3. The Legend of Sleepy Hollow | 3. Harris, Joel Chandler |
| 4. The Murders in the Rue Morgue | 4. Harte, Bret |
| 5. Innocents Abroad | 5. Irving, Washington |
| 6. The Outcasts of Poker Flat | 6. Poe, Edgar Allan |
| 7. The Raven | 7. Porter, Sidney |

The variations in form of the matching exercise are too numerous to be illustrated here. In the simple one-to-one matching type, the number of items in each column is the same and each item in one column is to be matched with one and only one item in the other. In the imperfect matching type, there are more items in one of the lists than in the other, some responses not being used at all (the first illustration is of this type). Again the pupil may be directed to use each term as a response as often as he thinks it is needed (the second illustration is of this type). The matching may be between names and names, names and descriptions, terms and definitions, questions and answers, causes and effects, dates and events, etc., with almost infinite variety. Diagrams, maps, and charts may be used. In a map exercise, for example, the pupil may be directed to write before each place name the number which locates that place on a numbered map. The matching may be multiple in character. For example, for each of the novels in a given list, the pupil may be directed to select from another list the name of the author and from

still another list the name of the principal character, writing the numbers corresponding to these names in the designated blanks opposite the name of the novel. The items to be matched need not necessarily be arranged in parallel columns. For example, a number of principles in physics might be given at the top of the test page, followed by a number of descriptions of situations in which one of these principles is applied. The pupil would then be directed to write in the blank before each description the number of the principle of which an application is illustrated. The possible variations in form and application, then, are limited only by the ingenuity of the test constructor.

Possibilities and Limitations:

1. The matching exercise has the advantage of compactness over the multiple-choice exercise, since the same set of alternate responses is used for a number of items, and also because it makes possible a greater simplicity in phrasing and structure through the possibility of indicating the general basis for matching, once and for all, in the directions to the exercise. It also makes practicable the use of a larger number of responses and therefore tends to reduce the guessing element.

2. The matching exercise suffers from the principal disadvantage that it is well adapted only to situations in which the items in at least one of the two lists to be matched consist only of single words, numbers, or very brief phrases, and is therefore not well adapted to testing for the acquisition of understanding of and ability to use relatively complex interpretative ideas. In general, the method of approach (see rule 13 for multiple-choice exercises) that it is necessary to employ in the matching exercise is the less desirable of the two.

GENERAL CONSIDERATIONS

3. The matching exercise is particularly likely to include irrelevant clues or cues to the correct response, unnoticed by the test constructor. It has perhaps been more frequently abused in this respect than any of the other types thus far considered. (See illustrations on pages 67 to 69.)

4. The matching exercise is particularly well adapted to testing in *who*, *what*, *when*, and *where* types of situations, or for naming and identifying abilities. Because of the ease with which it may be constructed and scored, it perhaps tends to be excessively used, resulting in undue emphasis upon non-interpretative materials in achievement testing.

Rules and Suggestions:

1. In general, do not include a large number of items in either of the columns to be matched. Twelve items in either column perhaps represents the maximum to be employed in most situations, with five or seven preferable. When longer lists are employed, too much of the pupil's time is used in "hunting" for the correct response.

2. In general, it is a good idea to include more responses than statements, in order to avoid the selection of the "last" response on an elimination basis.

3. It is often a good idea to allow certain responses to be used more than once, i.e., to constitute the correct response to more than one item in the opposite column, again in order to defeat elimination and guessing. This device forces the pupil to consider all responses in connection with each statement, whereas otherwise the student tends to cross out a response that has once been used and to give it no further consideration.

4. In every case, explain clearly the *basis* of the matching process that is involved. Never merely present two lists of items with the directions simply to "match," nor in-

clude more than one general basis for matching in the same exercise.

5. It is very important to include only homogeneous material in a single exercise. Each statement in the first column should have at least three plausible responses in the second column. If possible, all responses in the second column should constitute plausible responses to all of the items in the first column. Do not mix statements or responses highly dissimilar in character. For example, do not require the student to match men with events, dates with treaties, and events with locations, in the same exercise. The greater the degree of homogeneity, of course, the more certain and thorough must be the student's knowledge and insight. The test constructor should be definitely conscious of the degree of discrimination that he is requiring, and should not simply throw together a number of items to be matched without considering the problems that may be raised in the student's mind.

6. After an exercise has been constructed, invariably raise this question with reference to each item: What is the minimum amount of information or understanding that must be possessed by the student in order to eliminate each incorrect response or in order to select each correct response directly? (See pp. 73-78.)

7. Mechanical arrangement or assembly. In general, the column containing the longer statements should be on the left-hand side, and the shorter statements on the right-hand side. The student will then read the long statement first and search through the list of shorter statements, rather than read the short statement first and search through the list of long statements a number of times. If this plan is followed, the provision for the code number of the correct response should be made preceding the longer responses, i.e., at the

left-hand side of the page. (See illustration number 7, page 73.)

8. Where the "response" column consists of a number of words or names, arrange these in alphabetical order, in order to save the student's time in skimming the list. Dates similarly should be arranged in chronological order, numbers in numerical sequence, etc.

9. The matching exercise is not well adapted to testing over small units of subject matter because of the difficulty of finding a sufficient amount of homogeneous material to constitute a single exercise. In such cases it is ordinarily better to use the independent multiple-choice exercise.

10. Avoid the incomplete sentence structure in the construction of matching exercises. (See illustration number 2 on page 69.)

True-False Statement, or Alternate-Response Type

Form:

Usually a series of declarative statements, some of which are true and some false, each of which the student is to indicate as either true or false in the manner directed. (Illustrations will be found in the following pages.)

As was true of the preceding types, the alternate-response type of item appears in a wide variety of forms. The pupil may be directed to mark each statement as either *true* or *false*, or as *correct* or *incorrect*, or the statement may be in question form to be answered either *yes* or *no*. The pupil may respond by writing the word *true* or *false*, etc., in a blank; or by *underlining* the proper word, or by encircling the proper letter (T or F), or by writing + or -, or + or o, or Y or N, etc., in a blank, or by simply checking the true statements. The items may not always consist of statements or questions. They may, in a spelling test, be words correctly or incorrectly

THE CONSTRUCTION OF TESTS

spelled; they may, in an English test, be sentences which are grammatically correct or incorrect, or good or bad in sentence structure; etc. The items may be arranged in "cluster," all items in a single cluster relating to an introductory statement, paragraph, or diagram.

Possibilities and Limitations:

1. The true-false test has been perhaps the most widely used and the most seriously abused of all of the various types of objective test exercises.

2. The true-false test is perhaps the easiest of all types to construct, if quality is not considered. Unfortunately, in most instances, the method of construction has consisted essentially simply of skimming a textbook, selecting sentences more or less at random and using them almost intact as the bases for test items by simply including negatives or making minor changes in about half of the statements to make them false. As a consequence of this procedure, most of the true-false tests that have been built have tended to place a premium upon photographic recall of the printed pages of the textbook or of superficial and rote learning of unrelated and uniquely phrased statements.

3. The principal advantage of the true-false test lies in the ease with which it can be constructed and in the fact that a very wide sampling of individual elements can be tested in a short amount of time. Wherever the true-false test is made comparable in quality to other types, however, the amount of time saved in its construction is perhaps not significant. In most instances in which the true-false test has been used, it is perhaps true that the advantages of time saved and breadth of sampling have not been sufficient to make up for the lack of quality in the materials prepared.

4. Test technicians in general have steadily been losing con-

GENERAL CONSIDERATIONS

fidence in this technique, and at present it is employed in very few standardized achievement tests.

5. The true-false test appears to be particularly well adapted to those situations in which one wishes to test for the persistence of popular misconceptions or superstitious beliefs, where the suggestion of a correct response will make a multiple-choice item too obvious. It is also very well adapted to the situation in which it is impossible or extremely difficult to find enough plausible alternate responses to make a multiple-choice item, or in which there are only two *possible* responses, as in the following item: "In a lead-zinc cell, the lead plate is *positively* charged." Here there are obviously only two possible forms of the statement — "... is positively charged" and "... is negatively charged" — and the construction of a multiple-choice item would be impossible.

6. The true-false test has been characterized by serious ambiguities perhaps more frequently than any other type of test item. It would appear that this tendency toward ambiguity is inherent in the true-false statement.

7. The guessing element is more prominent in the true-false test than in any other type. For this reason particularly, an item analysis of the responses in a true-false test has relatively very little diagnostic value.

Rules and Suggestions:

1. The provision for the pupil's responses should be made as convenient as possible for both pupil and scorer. Directing the pupil to write the words *true* or *false* in blanks provided requires a needlessly large share of his time. Writing the letter T or F would be better, but these letters are likely to be difficult to distinguish in the student's handwriting. The provisions for responses in the illustrations below seem to be as good as any that have been suggested.

THE CONSTRUCTION OF TESTS

*Illustration No. 1*³

Directions: Encircle the T before all true statements and the F before all false statements. The first item has been marked correctly.

- (T) F 1. Like poles of magnets repel one another.
T F 2. The zinc can of the dry cell is the negative electrode.

Illustration No. 2

Directions: In the blank before each statement put a + if the statement is true and a o if the statement is false, as in the first two examples.

- ...o... 1. The electric currents in telephone lines are alternating currents.
...+... 2. Charges of static electricity of opposite sign attract one another.
..... 3. Rubbing cat's fur on an ebony rod produces a negative charge on the rod.

2. The number of true statements should be approximately equal to the number of false statements in the complete test.

3. Make the crucial element in the statement reasonably apparent to the pupil. Do not deliberately distract his attention from it.

- a. Do not bury a minor false element in a statement that is correct with respect to the elements emphasized. Particularly, do not include the false element in a qualifying

³ Items of this type can be readily scored by preparing a scoring key as follows: From one of the mimeographed copies of the test, cut from each page the strip containing the column of responses, i.e., the column of T—F's. On each strip, cut out the letter which is the correct response to each item, making each hole or "window" considerably larger than the letter cut out. This strip can then be superimposed on the pupil's responses for the appropriate page, and the number of correct responses on the page can then be told at a glance by simply counting the number of circles appearing through the "windows" of the key, thus eliminating the necessity of reading back and forth from the key to the pupil's paper when the ordinary parallel comparison method is employed.

GENERAL CONSIDERATIONS

phrase. Students in general tend to accept these as correct and to center their attention upon the major statement.

Illustration:

Bad: T F 1. The Sherman Anti-Trust Act, passed in 1870, declared combinations in restraint of trade illegal.

Better: T F 2. The Sherman Anti-Trust Act, which declared combinations in restraint of trade illegal, was passed in 1870.

In the first form of the illustrative statement, the pupil is likely to pass over the date as of no significance and to base his decision of truth or falsity upon the statement, "declared combinations in restraint of trade illegal." The form of the first statement tends to lead the pupil to disregard the crucial element, the form of the second statement forces him to consider it more carefully.

- b. In general, it is perhaps the best policy to have the crucial elements come at the end of the statement whenever possible. When the item consists of a number of sentences, make the last sentence the crucial statement and indicate to the student that the content of the introductory sentences may be accepted as correct. Make this clear to the pupil in the written or oral directions.
- c. In a "reason" or "because" item, include the false element, if at all, only in the "reason." Illustration:

Bad: T F 1. Grant's administration was marked by very little political scandal, because his own honesty was an incentive to those holding political offices.

Better: T F 2. Grant's administration was marked by serious political scandal because . . . (either correct or incorrect reason may be provided).

THE CONSTRUCTION OF TESTS

The pupil would naturally interpret the first statement as equivalent to the question "*Why* was Grant's administration marked by very little political scandal?" whereas the real purpose of the item is indicated by the question "*Was* Grant's administration marked by political scandal?" The first form of the statement is misleading and appears to be a deliberate "trap." Statements in which a deliberate attempt is made to distract the student's attention from the crucial element are analogous to "catch" questions in an essay examination. They tend to trip up the student whose knowledge is sound but who naturally ignores what should logically be minor or unimportant elements in the statement and who interprets a statement in the straightforward fashion characteristic of ordinary reading. Statements of this kind tend to test for the student's mental alertness or "test-wiseness" rather than for his knowledge or understanding of the subject matter involved.

4. Avoid the use of textbook language, particularly where figures of speech or unique phraseology are involved. This is important, not only because the use of textbook language encourages rote learning, but also because so many textbook statements become ambiguous when removed from context.

5. Avoid "specific determiners," i.e., external characteristics of an item that are usually associated with a true statement or a false statement. As was noted on page 72, words such as "all," "none," "always," may become specific determiners if used more frequently in false than in true statements. The point is *not* to avoid using such words, but to avoid using them only (or predominantly) in true or only in false statements. No words or characteristics are inherently specific determiners, they are made so only by the way in which they are used by the individual test constructor. The same word might be a specific determiner for a "true" response in the tests of one

GENERAL CONSIDERATIONS

teacher, for a "false" response in the tests of another, and might not be a determiner at all in still another teacher's tests.

6. In the two-response type of test, chance plays a larger part in determining the pupil's score than in any other type of test here considered, and is perhaps too large to be disregarded, even in the informal classroom testing situation. It is suggested that in all such tests the score be considered as the *difference* between the number of right answers and the number of wrong answers. (Score = number right — number wrong.) Items not attempted by the student are not considered directly, i.e., items omitted are not counted as wrong.

The literature of testing contains a large number of discussions of the "logic" underlying this correction-for-guessing formula. In the opinion of the writer, all such discussions are irrelevant and most of them are misleading. The justification for the use of this formula is empirical, not logical. Experiments have shown that scores on two-response tests obtained by means of this formula are in nearly all instances more valid and more reliable than scores on the same tests obtained by simply counting the number of right responses. As long as this is true, the use of the formula is justified, regardless of what the "logic" of the situation appears to be.

The experimental evidence indicates, furthermore, that even when scores on two-response tests are "corrected for guessing," the pupils should be cautioned against guessing. The directions to the pupil should contain the equivalent of the following: "If you feel you know nothing about an item, leave it alone. This does not mean that you are to answer only items about which you are dead certain. If you have any definite, rational basis for selecting an answer, do so, but do not guess wildly, since in scoring your paper a *deduction* will be made for each wrong answer you make."

7. Use quantitative rather than qualitative language wherever possible. Avoid the use of ambiguous terms such as "large," "many," "great," "small," "important," "unimportant," etc. Use direct comparisons wherever possible if a quantitative description is not possible.

QUESTIONS FOR DISCUSSION

1. Why does a topical analysis of the content of a test yield inadequate evidence of the proper distribution of items with relation to the objectives of instruction?
2. Why is it important that the table of specifications be multiple in character?
3. Why should one not attempt to determine in advance what type or types of items to employ or what proportion of items will be of each type in the construction of a general achievement test?
4. Find an actual test, either published or teacher-made, in which you believe the directions to the pupil are inadequate. Find also an actual illustration of poor typography or mechanical arrangement of a test, and indicate how each test might be improved in these respects.
5. Secure an actual distribution of test scores for a class of pupils and assign letter grades by the method described on pages 118-125.
6. What are the various defects which most frequently characterize objective test exercises of the simple recall type? For each of these defects, find at least three items, from actual tests, which illustrate it. Indicate how each item might be revised to eliminate the defect.
7. Select 10 important concepts from some field with which you are familiar, and construct a simple recall exercise testing for the pupil's understanding of each concept. For each item indicate the exact phraseology in which the rote learner would be most likely to have memorized the "correct answer," i.e., the textbook statement which contains or provides the basis for the right answer. Show why the mere memorization of the state-

GENERAL CONSIDERATIONS

ment is not likely to enable the rote learner to respond correctly to your item.

8. Find several illustrations of sentence or paragraph completion items that you consider to be overmutilated and therefore of the "puzzle" type. Why is it so difficult for the test constructor to avoid puzzle situations of this type? What is the best way to guard against them in test construction?
9. Find several illustrations of sentence or paragraph completion exercises that contain internal clues to the correct response. Draw specific attention to these clues and indicate how they might be eliminated.
10. Find three concrete illustrations of each of the defects which most frequently characterize the multiple-choice type of item. Show how you would improve each item by revision.
11. What is the principal advantage of the matching exercise over independent multiple-choice exercises testing for the same content? Why is it difficult to build good matching exercises over short units of content? Why is the matching exercise not well adapted to testing the pupil's understanding of or ability to use relatively complex or interpretative ideas?
12. Submit three examples, from published or teacher-made tests, of each of the defects commonly found in matching exercises. Suggest a revision for each example.
13. What are the principal reasons that the true-false type of exercise has been so seriously abused in test construction?
14. What defects are often found in items of the true-false type? Find several actual illustrations of each defect and suggest a revision which will correct the defect in each case.

PART II
EXAMINATIONS IN MAJOR SUBJECT
FIELDS

CHAPTER IV

EXAMINATIONS IN THE SOCIAL STUDIES

THIS chapter is intended to serve two purposes: (1) to suggest procedures for the construction of valid informal tests in the social studies, and (2) to suggest bases for the evaluation of published test materials in this field. It should be stressed also that this chapter is intended to supplement Chapters II and III, "The Theory of Test Construction" and "The Construction of Tests." As far as is practicable, pertinent content included in the earlier chapters has not been repeated. For this reason the three chapters in effect constitute a single unit dealing with testing in the social studies. To be wholly understood, the content of the present chapter must be considered in conjunction with that of Chapters II and III.

GENERAL ACHIEVEMENT TESTING

At two stages in the teaching of a given social studies course the teacher is especially interested in obtaining a reasonably accurate ranking of his pupils in order of their relative achievement in the field of subject matter involved. At the beginning of the course he needs this information in order to adapt instruction to the needs of individual pupils. At the end of the course he is interested in the comparative gains made by his pupils, as well as in their final ranking, for the purpose of determining promotions, assigning grades, and evaluating the effectiveness of his own instruction, and for use in the educational guidance of the pupils. For these purposes the teacher will use a general achievement type of test, either a standard-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

ized examination, such as those published by the Cooperative Test Service, or a test of his own construction. Whether he selects a published test or builds his own, he must appreciate the considerations underlying the construction of a general achievement examination.

Characteristics of a General Achievement Test

To build a valid and generally acceptable examination, the test constructor must have before him an authoritative and specific description of the subject matter to be tested. An achievement test is made up of specific items dealing with specific materials. No single test item can measure directly the attainment of an ultimate objective. Ultimate objectives merely represent convenient ways of describing collectively more immediate objectives, and the attainment of ultimate objectives can be measured only by measuring the attainment of the immediate objectives. It is of little value for the test builder to know, for instance, that an ultimate objective of the social studies is "to develop socially efficient citizens" and that an intermediate objective is "to develop in the pupil a reasoned understanding of social, economic, and political institutions, practices, and problems." He must base his test upon the specific items of information, the specific relationships, and the specific ideas, generalizations, etc., that constitute the accepted subject matter of instruction. The only authoritative descriptions of content thus far available to the test constructor are the textbooks and courses of study actually used in instruction. It is inevitable, therefore, that present standardized achievement examinations will follow closely the content of the best of present textbooks and courses of study.

Furthermore, it would be undesirable, for the purposes mentioned above, to build a standardized test the content of which would differ radically from the content of current instruction.

EXAMINATIONS IN THE SOCIAL STUDIES

If such an examination were administered widely, the rankings attained by various classes and schools would be influenced significantly by the degree of correspondence existing between the course as taught in a particular school and the content upon which the test was based. Clearly the most equitable basis for an examination that is to be used in a number of schools is the subject matter which is common to the best textbooks and courses of study in current use.

There is still another limitation on the elements of content which may serve as bases for the construction of individual items in a general achievement examination. This limitation grows directly out of the fact that the general achievement examination depends for its validity upon the degree to which it ranks pupils in the order of their true total achievement in the field of subject matter. This being the case, the test should include only items which discriminate rather sharply between pupils of superior and inferior total achievement. This consideration may serve to exclude from the test some elements of content of highest validity from a subject matter point of view.¹

In conclusion, then, it may be said that the classroom teacher will not find that the standardized general achievement examination parallels the content of instruction in his course as closely as can a test of his own construction. This limitation is not serious, however, since the discrepancy usually is not great. The special advantages of the standardized general achievement test are that it is made up of items carefully selected to conform to the dominant purpose of the test, and that it is provided with norms that make possible the comparison of pupil and class achievement from school to school.

¹ For a detailed discussion of the function of a general achievement examination, the range of difficulty of achievement test items, the discriminating power of test items, factors affecting discriminating power, etc., see pp. 26-65.

Steps in the Construction of a General Achievement Test

If the classroom teacher chooses to make his own general achievement examination he should follow the same procedure as does the constructor of standardized tests. First he must identify the elements of content, an understanding of which is necessary to the attainment of the ultimate objectives of the course, and must select a sampling of these elements to serve as the subject-matter basis for the construction of test items. He then proceeds to the actual construction of the individual test items. Of these steps the latter is by far the most important, since the crucial factor in determining the quality of the examination as a whole is the structural quality of the individual test items.

Selection of the Sampling. As practical guides in determining the content which properly belongs in the field to be tested the classroom teacher has the textbooks used, perhaps also a local course of study, and his own more or less detailed lesson plans. It is necessary, however, that he prepare a table of specifications indicating the relative emphasis which various phases of the course are to receive in the test. In this connection it should be recognized clearly that there are several independent bases upon which the test items may be classified. In a given history course the elements of content might be classified under chronological divisions, topical units, types of history — social, economic and political, or according to types of association — men and events, dates and events, events and locations, historical terms and meanings, cause and result relationships, etc. Some of these categories overlap, of course. Nevertheless it is mandatory that the content of the test be at least roughly checked against a number of such classifications. This necessity can best be demonstrated by an illustration of the general procedure which might be adopted in selecting a sampling of elements to serve as the content basis

EXAMINATIONS IN THE SOCIAL STUDIES

for test items to be included in a general achievement examination in high-school American history.

In the teaching of such a course, the teacher allots a certain number of days to each important period or major chronological division. Generally speaking, the time allotment corresponds closely to the teacher's conception of the comparative importance of the various periods. For convenience in discussion, it may be assumed that the following major divisions in American history receive the proportionate emphasis, i.e., percentage of total teaching time, indicated.

Major Divisions in American History

	Per Cent of Time
I. Discovery, exploration, and struggle for colonial supremacy (1492-1763)	10
II. Breaking away from the mother country (1763-1789)	15
III. Establishing the American nation (1789-1828)	10
IV. Territorial expansion and the slavery controversy (1828-1865)	15
V. The reconstruction period (1865-1876)	5
VI. Conflict between urban and agricultural civilization (1876-1898)	15
VII. Imperial expansion (1898-1914)	10
VIII. World War and after (1914-1936)	20

If the foregoing distribution of time is followed in teaching the course, it seems clear that the distribution of emphasis in the general achievement test also should follow it closely. In other words, in an examination of 200 items it would seem proper to include about twenty items over the first division, thirty over the second, etc. In the course of instruction the major chronological divisions, to be sure, are further "broken down" into topics. Thus, in the teaching of the first major

EXAMINATIONS IN MAJOR SUBJECT FIELDS

division the following organization might conceivably be followed:²

- I. Discovery, exploration, and struggle for colonial supremacy (1492-1763)
 - A. Conditions in Europe about 1500
 - B. Discovery and exploration
 - C. Colonization of America
 - D. The struggle with the French
 - 1. Causes of rivalry
 - a. In Europe
 - b. In the colonial field
 - 2. Early conflicts
 - 3. The French and Indian War
 - 4. The Peace of Paris
 - a. Terms
 - b. Significance

It has already been remarked that only about twenty items in the examination should be based on the foregoing content. Carried to its logical conclusion, this distribution of emphasis would probably limit the number of items dealing with "The struggle with the French" to about five, and permit one item to deal with either the terms or significance of the "Peace of Paris."

Even as meticulous a distribution of items in terms of major divisions as that here suggested by no means guarantees a satisfactory emphasis on the various phases and aspects of American history. For example, even though the items be distributed as suggested in the outline, a disproportionate emphasis might be placed upon political as contrasted to other aspects of history. Such lack of balance might be detected if the test items were further classified as fundamentally testing political, social, or economic aspects of history. This additional check, however, would not reveal whether a reasonable

² Subdivision "D" in this example is expanded in order to illustrate sampling.

EXAMINATIONS IN THE SOCIAL STUDIES

allotment of items had been made in terms of the major topics in which the content might be divided. This additional "cross check" might be made in terms of such a topical organization as the following:

1. Arts and Education
2. Finance
3. Foreign Affairs
4. Immigration
5. Industry and Labor
6. Nullification and Secession
7. Political Parties
8. Tariff Legislation
9. Territorial Expansion
10. Communication and Transportation

The reason for drawing up a table of specifications with these multiple classifications is to insure a representative sampling of elements of content from each of several points of view. In the building of items the classroom teacher probably will rely chiefly on such a classification as is suggested in the guidance outlines included in his lesson plans. This classification alone and of itself is neither adequate nor satisfactory. However, if the teacher is mindful of the multiple bases for classification while constructing the test items, it is likely that he will build a reasonably "well-balanced" test. In any event, when the tentative draft of the test has been completed the author should make a number of "cross checks" or "cross classifications" in order to insure satisfactory distribution of emphasis.

Construction of Individual Test Items. It has already been stated that the actual construction of test items is the most crucial factor in determining the quality of the general achievement examination. This is only another way of saying that if the individual test items are structurally or functionally defective, the examination as a whole is of poor quality regard-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

less of the care taken to insure a representative sampling in selecting the content bases of the items. In the hypothetical situation already described, the classroom teacher probably would want to include in his American history examination one item dealing with the Peace of Paris. He next must decide what particular element of related content should be tested.

Critics of the objective examination repeatedly assert that it does little but test the pupil's ability to recall or identify unrelated odds and ends of descriptive information. Unfortunately this criticism is justified, although it perhaps should be pointed out that the same charge could be directed, with even greater justice, against the traditional question and answer recitation in the social studies. In either case, there is likely to be a preponderance of such questions as:

When was the Peace of Paris concluded?

What war did the Peace of Paris bring to a close?

What nation profited most from the Peace of Paris?

What territory was acquired by Great Britain in 1763?

The author hastens to say that he does not imply that pupils ought not to learn such facts as are called for in the foregoing questions. What he wishes to stress is that the acquisition of descriptive facts by the pupil is not in itself a major goal of instruction. The desirable outcome is that the pupil acquire the ability to draw the generalizations and to grasp the relationships which may be based on the descriptive facts. ✓It is these broader understandings of interpretative ideas which should receive major emphasis both in teaching and testing.✓

✓Thus when the pupil studies about the Peace of Paris, two of the understandings he should acquire are: that the removal of the French and Indian menace caused the American colonies to feel less need for the protection of the mother country; and that this feeling of self-sufficiency caused the colonies to oppose the new imperial policy which Great Britain

EXAMINATIONS IN THE SOCIAL STUDIES

attempted to develop after 1763. In addition to these major understandings, there are others of only slightly less significance, centering around the following: Great Britain became the chief colonial power in the world and mistress of the seas; France ceased to be a major colonial rival of Great Britain; and Canada was thereby destined to develop under British social, economic, and political institutions. Finally, the pupil probably ought to know the time relationship of this peace treaty to relevant events preceding and following it. ✓

The desirable emphasis in testing is upon the measurement of real or reasoned understanding of significant information, relationships, generalizations, etc., as well as the ability to make interpretative use of these facts and ideas. This is in contrast to the unintentional but nevertheless actual emphasis upon verbal associations, upon the acquisition of isolated descriptive facts, and upon the memorization of unique or stereotyped textbook statements which so frequently characterizes both standardized and teacher-made tests.

Within the limits of this chapter it is impossible to illustrate in detail all of the different types of text exercises which may be used in social studies testing. It seems preferable rather to discuss in detail such exercises as have been found especially effective. This limitation of treatment possesses a further advantage in that it makes possible an incidental discussion of principles of test construction which must be followed if items are to measure reasoned understanding rather than rote learning.³

Matching Exercises

Matching types of exercises are adapted for testing understanding of technical vocabulary and of relationships between important events and personages, between events and loca-

³ For a discussion of the functioning content of test items, see pp. 66-81.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

tions, etc.⁴ 'One of the characteristics of a good matching exercise is that the responses from which the pupil is to make his selection must be sufficiently similar so that each will seem a plausible answer to any item in the exercise.' If this is not the case, the pupil may select the correct response on the basis of general information or superficial knowledge rather than on the basis of the specific understanding for which the item purports to test. Consider, for example, the following exercise:

- | | |
|---|--------------------------|
| ✓(4) The law that forbade slavery north of the Ohio | 1. Mason and Dixon Line |
| (1) A boundary between two colonies, that later became famous as the division between free and slave states | 2. Dred Scott Decision |
| (3) The fleet whose defeat in 1588 gave England control of the Atlantic | 3. Spanish Armada |
| | 4. Ordinance of 1787 |
| | 5. Missouri Compromise ✓ |

In selecting the answer to the first item, the pupil cannot fail to notice that only two of the five possible responses suggest a law. This suggestion is much the stronger in the fourth response, "Ordinance of 1787." A recognition of the general characteristics of the required response may thus enable the pupil to respond correctly to this item even though he does not possess the specific information called for, i.e., a knowledge of the provisions of the Ordinance of 1787. The second item is equally easy to answer. Three of the responses in no way suggest a boundary, and of the other two, one includes the word "line." The pupil will hardly fail to select "Mason and Dixon Line" in reaction to the statement beginning "A bound-

⁴The reader will find an evaluation of the possibilities and limitations of the commonly used types of objective test exercises, as well as specific rules and suggestions for their construction, in Chapter III, "The Construction of Tests." Matching exercises are discussed in pp. 147-152.

EXAMINATIONS IN THE SOCIAL STUDIES

ary..." Similarly, in view of the general characteristics of the foils (wrong responses), there is really only one possible answer to the last item. The third response is the only one which at all suggests a fleet. Furthermore, this response also includes the word "Spanish," which will suggest to the pupil the traditional enemy of England during the period indicated. In other words, the pupil, even though lacking the understanding called for in these questions, may be able to infer the correct answer in each case on the basis of only its general and external characteristics. ✓The only effective way to remedy this situation is to arrange the grouping so that all the items in an exercise will have the same general characteristics, i.e., so that all responses will deal with boundaries, or with the provisions of treaties and ordinances, etc.✓ The comparative-ness of such a homogeneous exercise may be inferred from a brief examination of the following example:

- | | |
|--|-------------------------|
| ✓(1) A boundary between two colonies, that later became famous as the division between free and slave states | 1. Mason and Dixon Line |
| | 2. Missouri River |
| (3) Marked the northern boundary of slave territory in the area immediately west of the Alleghenies | 3. Ohio River |
| | 4. 49° |
| (5) Was intended to be the boundary between free and slave territory in the Louisiana Purchase | 5. 36° 30' |
| | ✓ |

Pupils who select the correct responses in exercises of this character are much more likely to possess a genuine understanding of the terms tested than are those who can make the correct associations only in less homogeneous exercises. Homogeneous grouping of materials is, of course, just as im-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

portant in tests informally constructed for classroom use as it is in formal standardized tests.

In constructing matching exercises of highest quality it is not enough merely to strive for homogeneous grouping of test items. ✓ It also is extremely important to avoid the use of textbook language and stereotyped phraseology.⁵ ✓

To illustrate the manner in which these desirable characteristics may be incorporated into a matching exercise, consider briefly the following example:

- | | |
|--|--------------------|
| ✓(4) His revolt against the Church probably would have been unsuccessful had the Emperor not been engaged in foreign warfare | 1. Henry VIII |
| (1) His conflict with the Church netted him great economic advantages | 2. John Huss |
| (3) He was instrumental in reclaiming a large part of Germany for Catholicism | 3. Ignatius Loyola |
| | 4. Martin Luther |
| | 5. St. Dominic ✓ |

✓ Each of the statements in this exercise has been worded to avoid textbook language, and all of the responses are homogeneous in their general characteristics, i.e., each of the men listed had important direct relationships with the Church. ✓ In responding to the first statement, the pupil not only must recognize which man revolted against the Church, since there are three such men listed — Henry VIII, Huss, and Luther — but also must be able to infer that, of the three, Luther is the only one whose revolt both was successful and logically should have been opposed by the Emperor. Similarly, in relation to the second statement, three of the responses have the desired

⁵ See discussion of rote learning with special reference to the new-type test, pp. 81-96.

EXAMINATIONS IN THE SOCIAL STUDIES

general characteristics, since three of the men listed were in conflict with the Church. The pupil who knows that neither Huss nor Luther was primarily influenced by economic motives would have no difficulty in selecting Henry VIII. It is clear, however, that pupils who possess only superficial information, or who do not have a thorough understanding of the rôle played by each of the three men in the Reformation movement, could hardly guess the correct response in this exercise.

Multiple-Choice Exercises

Whereas the matching type of exercise is well adapted to testing the pupil's ability to associate worthwhile meanings with geographical terms, technical phrases, or events and institutions which may be identified by name, the multiple-choice or best-answer type of exercise is most effective for testing broader implications, particularly of things which cannot be thus identified. Multiple-choice exercises can be used effectively in an attempt to measure such things as cause and effect relationships, factors underlying important developments, consequences of certain actions, and the significance of certain political, social, and economic situations and practices. Thus in many respects multiple-choice exercises may constitute the most significant section of the social studies examination.⁶

The following examples taken from recent Iowa Every-Pupil tests may serve to indicate specifically the type of items which may be developed.⁷

⁶ The reader will find a detailed account of the possibilities and limitations of the multiple-choice exercise, as well as specific suggestions for its effective construction, on pp. 136-147.

⁷ Russell C. Ross, *An Analysis of the Data Secured from the Iowa Academic Test in World History*. Unpublished M.A. Thesis, State University of Iowa, August, 1932, p. 53.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

- ✓(3) A major policy upon which the Italian Fascists and the Russian Bolsheviks are in agreement is that of
1. Establishing a dictatorship of the laboring class (26%)⁸
 2. Abolishing the private ownership of capital (21%)
 3. Denouncing the doctrines of democratic government (16%)
 4. Opposing imperialism as a national policy (19%) ✓
(Item omitted by 18 per cent of the pupils.)

This item illustrates the possibility of phrasing responses in such a way that each may appear plausible to the pupil who has only a superficial knowledge of history. It is commonly known that Bolshevik leaders advocate the policies indicated in the first two responses. Even the inferior student is likely to have this information. To the pupil not well informed about Fascist policies, therefore, the first two responses will appear particularly plausible. The fourth response may appear reasonable to any pupil who does not know the meaning of "imperialism," or who interprets it, as many may, as somewhat synonymous with or frequently associated with "royalism." The well-informed pupil, however, who knows that the Soviet government ostensibly disclaims imperialistic ambitions, also should know that Mussolini repeatedly has asserted the right of Italy to increased colonial territories as an outlet for a rapidly expanding population. From this knowledge he should readily be able to eliminate the fourth response.

An analysis of pupil responses to this test item reveals that

⁸ The figures in parentheses indicate what per cent of the 7528 pupils tested in world history in the 1932 Iowa Every-Pupil program selected each response. The percentage of pupils selecting each response is similarly indicated for other test exercises included in this chapter.

EXAMINATIONS IN THE SOCIAL STUDIES

each of the three foils was chosen more frequently than the correct response. The possibility of making incorrect responses appear extremely plausible causes an item of this type to call for a more thorough and certain understanding than is required in the case where, as in an essay question, the pupil is asked simply to recall the correct answer without any suggestions. In the recall type of test the pupil is likely to write down the first answer that occurs to him. Relying as he does upon memory rather than upon understanding, he is likely to record a "pat" statement which may be correct, but which he does not fully comprehend. In the multiple-choice exercise, the same idea, differently phrased, will often appear incorrect, particularly when the pupil's attention is attracted by other plausible answers which ordinarily would not have occurred to him.

The next item illustrates an attempt to measure the pupil's understanding of one of the causes of an important action.

- ✓(3) The adoption of the Constitution was generally opposed by
- | | |
|----------------------------------|-------|
| 1. Prosperous merchants | (18%) |
| 2. Wealthy manufacturers | (21%) |
| 3. Small farmers | (21%) |
| ✓4. Holders of public securities | (38%) |
- (Item omitted by 2 per cent of the pupils.)

The question, "What classes opposed the adoption of the Constitution?" is not one to which a "pat" answer is likely to be found in the pupil's textbook, and therefore it will be necessary for him to do some inferential thinking to answer it correctly. It is nevertheless a fair question, since the facts needed should be in the possession of all well-taught pupils. ✓ Most texts stress the economic implications of the adoption of the Constitution, and emphasize the connection between a strong central government and financial stability. To the pupil

with a reasoned understanding of these implications and relationships, it should be clear that holders of public securities, prosperous merchants, and wealthy manufacturers would have strong reasons to advocate, rather than to oppose, the adoption of the Constitution, and that only the small farmers, traditionally members of the debtor class, would lose nothing if the government remained weak and financially unstable.

From the analysis of the responses to this item, however, it appears that only one out of five (21 per cent) of the 7800 Iowa pupils tested in American history in the 1932 Every-Pupil program had attained this degree of understanding.⁹

Chronology Exercises

✓ This type of exercise will be discussed in considerable detail in order to illustrate specifically the need for developing testing procedures which really measure what the exercise purports to measure, and to illustrate further the extreme importance of careful phrasing of items. Consider briefly the following examples:

Type A

Directions: Match each event with the proper date.

- | | |
|---------------------------------------|---------|
| (2) Sherman Anti-Trust Act | 1. 1876 |
| | 2. 1890 |
| (5) Washington Disarmament Conference | 3. 1905 |
| | 4. 1914 |
| (4) Opening of the Panama Canal | 5. 1922 |

Type B

Directions: In the following exercise, an historical event is described or suggested by each of the statements in the left-hand column. Match each statement with the time period in the right-hand column in which the event occurred.

⁹ Loren Bane. *Analysis of the Every-Pupil Test in United States History of the 1932 Iowa Academic Contest*. Unpublished M.A. Thesis, State University of Iowa, August, 1932, p. 32.

EXAMINATIONS IN THE SOCIAL STUDIES

- | | |
|--|------------------------------|
| (2) A federal statute <i>was passed</i> which later became the basis for a suit to dissolve a merger of the Great Northern, the Northern Pacific, and the Chicago, Burlington and Quincy railroads | 1. 1876-1885
2. 1886-1895 |
| (5) At an international conference, the United States secretary of state proposed that the building of first-class battleships be discontinued for ten years | 3. 1896-1905
4. 1906-1915 |
| (4) The completion of a great engineering project reduced, by about two-thirds, the distance by boat from New York to San Francisco | 5. 1916-1925 |

Type C

✓Directions: In the following exercises, the four events in the right-hand column are arranged in the order in which they occurred. Each of the *numbers* in the right-hand column, therefore, corresponds to a *time interval*. Interval 1 is that preceding the first event, interval 2 is that between the first and second events, and interval 5 is that following the last event. Each of the statements in the left-hand column describes or suggests an historical event. For each of these statements, indicate the time interval in which the implied event occurred by writing the appropriate number in the parentheses preceding the statement.

- | | |
|--|---|
| (2) A federal statute <i>was passed</i> which later became the basis for a suit to dissolve a merger of the Great Northern, the Northern Pacific, and the Chicago, Burlington and Quincy railroads | — 1 —
Civil War
— 2 —
Spanish-American War |
| (4) At an international conference, the United States Secretary of State proposed that the building of first-class battleships be discontinued for ten years | — 3 —
Close of the World War
— 4 — |
| (3) The completion of a great engineering project reduced, by about two-thirds, the distance by boat from New York to San Francisco | London Naval Conference
— 5 — |

✓The first example (Type A) illustrates a type of exercise which has frequently been employed in history testing. All

EXAMINATIONS IN MAJOR SUBJECT FIELDS

that it requires of the pupil is a verbal association between a date and a name. For instance, if the pupil has memorized "Sherman Anti-Trust Act — 1890" and "Washington Disarmament Conference — 1922," etc., he can respond correctly to the items in this exercise even though he knows nothing about the provisions and the scope of the Sherman Anti-Trust Act, about the results achieved by the Washington Conference, or about the significance of the opening of the Panama Canal.

The second example (Type B) illustrates a type of exercise which is far superior to the first. Each event is identified by a brief description of an important aspect, consequence, or implication of that event, rather than by its name only. The exercise now calls for some understanding of the nature or significance of each of the events in question, rather than for only a verbal association which may have been learned by rote. For example, the pupil must know enough about the Washington Disarmament Conference to understand that it is implied in the statement, "... the building of first-class battleships be discontinued for ten years," before his verbal association between name and date can function in selecting the proper time interval.

The phrasing of items of this type is particularly important. Consider, for example, the following descriptions of a single event, any one of which may be employed in a Type B exercise.

At the Washington Conference, Secretary of State Hughes proposed a "naval holiday" in the building of battleships.

At an international conference, Secretary of State Hughes proposed a "naval holiday" in the building of battleships.

At an international conference, the secretary of state of the United States proposed a "naval holiday" in the building of battleships.

At an international conference, the secretary of state of the United States proposed that the building of first-class battleships be discontinued for ten years.

A correct response could be made to the first item by any pupil who had memorized "Washington Conference — 1922," even though the rest of the statement meant nothing to him. The second phrasing is better, but the pupil could place the event in the proper time interval if he knew only at what time Hughes was secretary of state, even though he knew nothing of the Washington Conference. The third statement eliminates this "clue," but still contains the catch phrase "naval holiday," which the pupil might have associated with the proper date and event without appreciating much of its meaning. The pupil who recognizes the event as described by the last item, and who can thereupon place it in the proper time interval, is likely to have a better understanding of it than one who could answer any of the three preceding items but not the last.

In phrasing the responses to items, an attempt must be made to eliminate "clues" irrelevant to the point in question, as well as to avoid the use of "pat" or "catch" phrases that the pupil may have memorized without appreciating their full meaning. The phrasing employed, however, must be free from ambiguity, i.e., it should suggest the event in question, and not any other. Finally, it should suggest the event in such a way that it is fair to expect the pupil to recognize the association, i.e., it should deal with a significant aspect or consequence of the event.

Type B, then, represents a superior test exercise. It still suffers, however, from the limitation that it does not require the pupil to relate events directly to one another, either logically or chronologically. ✓Type C, on the other hand, exhibits the virtues of the Type B exercise, and in addition does require the pupil to relate each event directly to other events. In this exercise, for example, the pupil must not only have some appreciation of the significance of the Washington Dis-

armament Conference, but must also recognize its relationship to earlier and later happenings.✓ Even though he may not be able to recall the exact year (which, after all, is relatively unimportant) of the conference, he should appreciate why the curtailment of competition in the building of expensive battleships was seriously agitated for the first time following the World War. He will then be able to infer that the Washington Conference must have been held following that world conflict. At the same time, the pupil should recognize that the London Naval Conference was called for the purpose of completing the unfinished work of the Washington Conference, and hence, certainly must have followed it. Similarly, if the pupil understands the logical or cause and effect relationships between the other events listed, he will be able to respond correctly to the exercise without necessarily resorting to a direct recall of exact dates. It should be clear, then, that the ✓Type C exercise is far more than a test of chronology or date-event relationships alone; that it calls into play much of the same sort of reasoning or understanding that is involved in all aspects of the learning of history.✓

Up to this point the writer has concerned himself only with three types of objective test exercises. He does not imply, however, that other types are lacking in merit. Generally speaking, the test builder should construct those types of test items which he feels he can best build in conformity with the characteristics of good test items as here discussed. Further adaptations of objective exercises already mentioned, as well as suggestions for other types of test exercises, are found in the subsequent pages of this chapter devoted to discussions of diagnostic testing, the use of the essay question, and the testing of attitudes.¹⁰

¹⁰ See Chapter III for further detailed suggestions for the construction of various types of objective test items.

All the Reasoning Necessary to Respond Correctly to Test Items Need not be Done During the Examination Period

The various sections of a general achievement test should emphasize to a high degree the principle that a proper study of social sciences should result, not only in the acquisition of unrelated facts or of information concerning isolated events, but in an appreciation and a reasoned understanding of the broad implications and consequences of a whole series of related events, of major movements, and of important institutions and practices. It should not be inferred, however, that all of the reasoning required for a correct response to each of the test items must be done by the pupil during the test period. Good teaching will have led the pupil to consider the basic materials upon which the test items are based. The pupil who already has thought through these problems will, to a large extent, need only to recall or repeat previous reasoning. The point is that, while the items will in most cases call for reasoned understanding rather than rote learning, most of the reasoning should have been done during the course of instruction and prior to the examination period, rather than during it. The pupil who has done no such thinking previous to the test period undoubtedly will be unable to respond correctly to all of the test items, no matter how well he may be posted on historical facts.

The Relations Between Teaching and Testing

It is perhaps not amiss to point out incidentally that the principles brought out in connection with test construction must also, to a considerable extent, characterize effective instruction. One of the discouraging aspects of present social studies teaching is that it tends to produce "human encyclopedias," pupils who can glibly furnish pat answers to pat questions of the "who, what, when and where" variety, but who

have failed to acquire any real understanding of history, economics or government. This excessive emphasis upon the acquisition of descriptive information is in part the cause, in part the result, of the practice of devoting a large part of the class period to a question and answer type of recitation in which the only contribution made by the pupil is a glib repetition of the "pat" phraseology of the textbook. Under this procedure a pupil may give a perfect recitation without knowing what he is talking about. A pupil thus taught often has very vague and even incorrect notions of terms and phrases used in the course of routine recitations. He may well memorize without understanding such statements as, "The chief cause of the American Revolution was England's new imperial policy," and "The chief purpose of the Federal Reserve System was to make the currency more elastic."

The Use of Tests Throughout Instruction

In the course of instruction over a given topic the classroom teacher makes use of objective drill exercises of all kinds, mastery tests, unit examinations, etc. These informal examinations in a sense serve a double function. In many cases the teacher makes use of pupil scores on the tests as a basis for assigning marks, thus emphasizing the characteristic of the general achievement examination; namely, the ranking of pupils in order of achievement in a given field of subject matter. He also makes use of informal tests to discover deficiencies in pupil achievement and to discover areas where reteaching and remedial teaching are necessary.

On the whole, these informal tests should consist of items of the same type and quality as those already discussed. To the extent that the fundamental purpose of the informal examination is to discover whether pupils have mastered certain concepts, relationships, etc., and to identify the pupils who

EXAMINATIONS IN THE SOCIAL STUDIES

have not done so, it is not necessary to be concerned about the distribution of item difficulty. In other words, if the chief purpose of the informal test is to discover deficiencies in pupil achievement, it will differ from the general achievement type of test in two particulars: The teacher will be less interested in the total score on the test and more interested in the responses to individual items; and the informal test usually will deal with a more restricted field of subject matter and consequently will make use of a more intensive sampling of elements in that field.¹¹

If the informal examination is to serve as a basis for determining the need of remedial instruction, it is important that the individual test items be sufficiently specific and searching to have diagnostic value. Recently a detailed examination over certain phases of high-school American history was administered to 422 pupils who had completed the work in question.¹² The examination was made up of multiple-choice items so constructed as to provide a "cross check" on the pupil's grasp of information bearing on important concepts. The following pair of items is typical:

(1) The national Bill of Rights prevents

1. Congress from making laws restricting freedom of the press
2. The states from limiting the freedom of the press
3. The press from criticizing the government
4. The President from censoring the press
5. The courts from suppressing newspapers that criticize the government

¹¹ For a discussion of the characteristics of a diagnostic test, see pp. 20-26.

¹² Hiram J. Eininger, *Pupil Information Bearing on Important Topics in American History, 1789-1798*. Unpublished M.A. Thesis, University of Iowa, August, 1933, pp. 49-51.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

- (5) Freedom of the press, as provided for in the national Bill of Rights, means that newspapers have the right to
 1. Comment maliciously on events in the private lives of citizens without fear of being sued
 2. Print anything whatsoever without restriction
 3. Send their editors to Washington to lobby for legislation
 4. Send reporters to all sessions of the legislative, executive and judicial branches of the government
 5. Criticize the policies of the government without fear of suspension of publication.

Eighty per cent of the total group (388 pupils) responded in such a manner as to indicate that they knew the Bill of Rights restrains Congress from passing laws restricting the freedom of the press. Yet 220 pupils out of the 422 selected responses in the second item indicating that they did not understand the meaning of the term, "freedom of the press." No less than 126 pupils indicated that they believed "freedom of the press" meant that newspapers had the right to "print anything whatsoever without restriction." Perhaps the most important discovery growing out of this particular analysis was that only 169 pupils out of the 338, exactly one-half, who responded correctly to the first item were able to select the correct explanation of the term "freedom of the press" in the second.

Evidence of this kind could be multiplied indefinitely to indicate the need for more exact and searching examination into the degree of genuine understanding and insight attained by pupils. The evidence also calls attention to the state of half learning, or worse, which characterizes so much of social studies instruction, and points out the crying need for remedial teaching of elements of content which pupils supposedly have mastered.

EXAMINATIONS IN THE SOCIAL STUDIES

Summary: The Criteria of the Achievement Examination

The purpose of the discussion to this point has been to provide classroom teachers of the social studies with standards in terms of which they may judge published standardized examinations and plan tests of their own construction. It seems desirable, therefore, to phrase a series of crucial questions which the teacher must consider in preparing his own examinations.¹³

- ✓ 1. Is the sampling satisfactory?
 - a. Have tables of specifications been drawn up to provide a number of independent classifications of content?
 - b. Does the suggested emphasis insure balanced emphasis on all phases of content?
- 2. Are the individual test items satisfactory?
 - a. Is each element of content tested for in the most effective manner?
 - b. Is the item phrased so as to be free of irrelevant clues and ambiguities?
 - c. Does the item test for reasoned understanding of content by avoiding stereotyped phraseology and textbook language?
 - d. Is the element tested for in the item significant in the light of the accepted purposes of instruction? ✓

TESTING OF WORK SKILLS

Heretofore this chapter has concerned itself with examinations which are intended primarily to measure the pupil's reasoned understanding of content. Similarly, in the classroom the major emphasis in teaching is upon the mastery of content. There is, however, in both teaching and testing, a distinct need for increased attention to *how* the pupil learns.

¹³ The reader should carefully consider the detailed rules and suggestions for objective test construction given on pp. 108-125.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

For example, if the pupil cannot read a given assignment meaningfully, he will not be able to master the significant elements of content found in that assignment. In such a situation the teacher cannot be concerned solely with mastery of content. Rather he must find out first whether the pupil can comprehend printed material. If the pupil is unable to do so, remedial procedures must be devised to develop his ability along this line.

The measurement of work skills has been generally neglected in informal classroom testing. It is the purpose of this discussion to call attention to certain important skills in the social studies, and to suggest procedures for their measurement. Limitations of space restrict the treatment to the following:

1. Reading to locate information
2. Summarizing
3. Outlining
4. Interpreting cartoons, graphs, charts, tables, etc.
5. Use of books
6. Map reading

1. Reading to Locate Information

The typical teaching procedure in the social studies classroom consists primarily of: the assignment of lessons in textbooks; and the oral quizzing of pupils over lesson assignments. This procedure is singularly inefficient. In the first place, the tendency is to assign substantially the same materials to all pupils regardless of their range in ability in silent reading comprehension. Secondly, the oral quiz centers on only one pupil at a time, and hence, is little calculated to hold the active interest of the entire group. To the extent that the teacher feels a need for holding all pupils responsible for the same materials, it is mandatory that he develop a technique which will require the active interest of all pupils, and which will

EXAMINATIONS IN THE SOCIAL STUDIES

secure comparable facts concerning the mastery of content from all pupils in the shortest possible time.

How this may be done will be discussed in connection with a specific illustration. Suppose, for example, the pupil is given the following materials to read:

The Second Bank of the United States

Par. 1 The first important act of the Fourteenth Congress was to create the Second Bank of the United States. In the days of Hamilton the Democratic-Republicans had opposed the bank on the grounds that the Constitution made no specific mention of such an institution. In 1811 their votes had killed the attempt to renew its charter. Yet a few years later these early opponents voted for the new bank. The following conditions largely account for their changed point of view.

Par. 2 Following the expiration of the charter of Hamilton's Bank in 1811 its place was taken by a number of institutions chartered by the several states. These banks varied greatly in strength and soundness. In five years they increased in number from 85 to 246 and their note issues mounted from \$50,000,000 to about double that amount. It was expected that the notes of the new state banks would circulate freely at par just as had those of the Bank of the United States.

Par. 3 This expectation was never realized. The notes issued by Hamilton's Bank were secured by coin, and hence circulated everywhere at their face value. The notes of the new banks had little or no gold back of them. Since they could not be turned into coin at the option of the holder, they fluctuated in value like the paper money issued during the Revolution. The seriousness of the inflation became apparent in the course of the Second War with Great Britain. The news of the capture of Washington by the British creating widespread alarm and unrest; all state banks outside of Massachusetts were compelled to refuse to redeem any of their notes in coin.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

Par. 4 The financial position of the national government itself was precarious. It had been impossible to finance the war from current taxation. Between 1812 and 1816 bond issues to a total face value of \$80,000,000 were floated. Yet the returns measured in terms of specie were disappointing — only about \$34,000,000. By these operations the national debt reached the staggering total of about \$125,000,000.

Par. 5 In this strait the Democratic-Republicans were virtually compelled to reverse the stand taken in 1811 and to charter the Second Bank of the United States. In most respects the new bank was like its predecessor. Its capital stock was \$35,000,000 as against \$10,000,000 formerly, one-fifth of which was in both cases subscribed by the government. It was to act, like the First Bank, as a depository for the funds of the federal government, and of course retained the right to loan out this money at interest. Its notes, well secured by coin, were to circulate as money. Nevertheless, the new bank was handicapped in one important respect. The state banks, jealous for their rights, were certain to do everything possible to harass their large and powerful rival.

If the teacher wishes to discover whether or not the pupils understand the foregoing selection he may ask them to answer a series of questions. To the extent that the questions are genuinely thought-provoking, i.e., do not make use of textbook language nor permit pat answers in terms of the exact language of the text, this line of attack has merit. Of course, if comparable information is to be secured from all the members of the class, the exercise must be written. The following questions, for example, might be assigned over paragraph 4 of the selection:

1. Was the federal "budget" balanced? Explain.
2. Were the government bonds marketed at above or below par? Explain.
3. What was the condition of government credit? Explain.

The answers to these questions are clearly implied in the fourth paragraph. Unless the pupil senses these implications,

EXAMINATIONS IN THE SOCIAL STUDIES

and in so doing he must understand such terms as "balanced budget," "above or below par," and "credit," the selection can have little meaning for him. The inexperienced questioner might try to get at the sense of the paragraph by asking, "What was the financial position of the government?" The pupils examined doubtlessly would agree that "precarious" was the correct answer. This somewhat exaggerated illustration indicates how meaningless pat answers to leading questions may be.

Pupil answers to questions such as the three listed will vary in quality from those unquestionably wrong to those unquestionably right. The subjective evaluation of the "borderline" responses doubtless tends to invalidate this type of exercise as a testing device. The grading of pupil responses obviously will be time-consuming, which fact is an even more serious objection. A remedy in this case is to convert the questions into the true-false type of objective items, as follows:

1. The federal "budget" was balanced in the period 1812-1816.
2. During this period government bonds were marketed at below par.
3. The credit of the federal government was good.

When the questions are adapted to true-false scoring, the pupil is not held responsible for any statement of why he believes what he believes. This desirable characteristic of the essay question may be retained to a considerable extent if a multiple-response exercise is used.

Which two of the following statements explain why the finances of the national government were in a bad condition?

1. Taxes were greatly increased during the war.
2. The government went heavily into debt.
3. People usually refused to pay taxes.
4. Newly issued government bonds sold at a discount.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

The multiple-choice type of exercise seems best adapted for the purpose discussed. It is particularly important here that the phrasing used differ from that of the selection upon which it is based so as to avoid the possibility of the pupil's selecting the correct response merely on the basis of recognition of similar or identical words or phrases.

2. Summarizing

The ability to summarize adequately and fairly what has been read is another important factor in reading comprehension. Social studies teachers commonly place a great emphasis on summarizing as a learning procedure. There are a number of different procedures which may be used to test this work skill. The first which comes to mind is to have the pupil give orally or in writing a summary of the selection read. This procedure, however, is better suited to drill than to testing purposes. A device which may be used to test the pupil's grasp of the total meaning of a selection is to ask him to indicate which of several summaries he feels is most nearly correct and adequate. The following is an example:

Which of the following is the most nearly correct and adequate summary of the selection, "The Second Bank of the United States"?

No. 1 In 1811 the Democratic-Republicans had favored the chartering of a number of state banks to take the place of the First Bank of the United States. They changed their stand five years later and chartered the Second Bank of the United States, an institution similar in function and organization to its predecessor. This was a necessary step because the state banks were refusing to pay cash for government bonds sold to finance the war. They also were making great profits by introducing the gold standard.

No. 2 When the First Bank of the United States passed out of existence, a number of state banks attempted to take its place. These issued bank notes which had little gold back

EXAMINATIONS IN THE SOCIAL STUDIES

of them. Consequently the notes were not always acceptable at face value. At last most of the banks refused to exchange them for coin. During this same time the government was having difficulty in marketing its bond issues. The Democratic-Republicans had little choice but to charter the Second Bank, an institution in most respects similar to the one they had opposed in 1811.

- No. 3 The Second Bank of the United States was the direct successor of the First Bank. The failure of state banks to buy government bonds had made it very difficult to finance the War of 1812. Another reason for chartering the Second Bank was that the Democratic-Republicans did not want the state banks to make all the profit from going off the gold standard. Naturally the state banks hated their large and powerful rival.

It is possible also to develop exercises of the multiple-response type to test whether or not the pupil can recognize the most important elements to include in a summary, and also whether or not he can recognize the central theme of a selection.

Which two of the following statements are most necessary to include in a brief summary of the selection, "The Second Bank of the United States"?

1. The Democratic-Republicans had no fixed financial policies.
2. The number of state banks increased in number from 85 to 246.
3. The bank notes issued by the state banks did not circulate at par and were not readily redeemable in gold.
4. The capital stock of the Second Bank of the United States was \$35,000,000 — one-fifth of which was subscribed by the government.
5. Paper money is not as good money as gold coin.
6. The national government had difficulty in profitably disposing of bond issues.
7. The capture of Washington by the British created widespread alarm and unrest.
8. The War of 1812 was financed entirely through current taxation.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

In the foregoing exercise the pupil is called upon to make a careful selection. Some of the statements included are false, some are irrelevant, and some are comparatively less important than others. The pupil is called upon to reject all but the two most significant statements. Clearly, the items selected must be those most necessary to a fair and adequate summary of the topic read.

The problem confronting the pupil in the next exercise is somewhat similar to that just discussed. In this case the pupil may feel that the selection he has read can be said to throw light on several of the themes suggested. His assignment, however, is to select the one statement which best describes the probable purpose of the author in writing the selection in question.

Which of the following do you think best describes the main purpose of the author in writing the selection, "The Second Bank of the United States"?

1. To explain why a gold standard is necessary
2. To describe the functions of the National Bank
3. To explain why state banks inflate the currency
4. To show how the Democratic-Republicans happened to come into power
5. To explain why a National Bank was needed in 1816

3. Outlining

Basically, outlining and summarizing are alike in that each involves: organizing the content under logical headings, discarding irrelevant elements of content, and emphasizing the major points of the content. Indeed, practice in outlining is perhaps the best approach to the development of skill in summarizing.

One test of a pupil's ability to outline is to have him submit an outline which he has developed independently over a given topic. This procedure has value since it recognizes and de-

velops individual initiative. Grading of such work is somewhat difficult, both because of the amount of time needed for adequate evaluation and because objectivity in the grading is difficult if not impossible to achieve. Some teachers assign pupils the task of rearranging a scrambled outline. Another procedure is to give the student an outline with the major headings written in and the number of sub-points merely indicated. The task is for the pupil to sense the proper statements to be written as sub-points. This device may be illustrated in terms of the following outline developed over paragraphs 2, 3, and 4 of the selection "The Second Bank of the United States."

Why the Democratic-Republican Party Changed Its Stand on the Bank Question

- I. The general economic situation in the United States following the expiration of the charter of the First Bank
 - A. The chartering of state banks
 - 1.
 - 2.
 - 3.
 - B. The comparative inefficiency of the state banks
 - 1.
 - 2.
- II. The unfavorable financial situation of the government
 - A. The War of 1812 necessitated heavy expenditures
 - 1.
 - 2.
 - B. The national debt mounted to a total of \$125,000,000

In this case the pupil is given the statements prefaced by Roman numerals and capital letters and is expected to write in the statements which he feels should logically follow the Arabic numerals. In this type of exercise the grading becomes reasonably objective, provided that the selection is so logically

EXAMINATIONS IN MAJOR SUBJECT FIELDS

written that there can be no reasonable doubt as to what are the sub-points to be included.

Through a use of the multiple-response type of exercise, it is possible to assign a definite score to the pupil's ability to recognize certain factors involved in the organization of ideas. Thus the following question might be based on paragraphs 2, 3, and 4 of the selection.

If the heading for an outline over paragraphs 2, 3 and 4 is, "Why the Democratic-Republican Party Changed Its Stand on the Bank Question," which two of the following statements would you include as major sub-topics?

1. The War of 1812 necessitated heavy expenditures
2. The general economic situation in the United States, 1812-1816
3. The unfavorable financial situation of the government
4. The comparative inefficiency of the state banks

If ingeniously constructed, the latter type of exercise can be made to test practically all of the abilities involved in the independent construction of a complete outline. At the same time, it should be recognized that this is primarily a measuring device and is not the most effective way of developing skill in the organization of ideas.

4. Interpreting Cartoons, Graphs, Charts, Tables, etc.'

In his daily reading of newspapers and magazines the adult is called upon to interpret cartoons, graphs, charts, tables, etc. These are included in the press on the assumption that they are easier to interpret than the same data presented in written discussion. The trend in social studies textbooks is toward increased emphasis of such aids to understanding. Clearly, it is essential to discover to what extent the pupil can read and interpret materials of this kind. There is no justification for what appears to be a rather common assumption; namely, that no special instruction in the interpretation of cartoons, graphs, charts, etc., is needed.

EXAMINATIONS IN THE SOCIAL STUDIES

There are a number of ways in which the teacher may gauge the ability of the pupil along this line. He may post a cartoon on the bulletin board, afford every pupil an opportunity to study it, and then ask that interpretative reactions be submitted in writing. Conditions affecting results on the examination may be further controlled if the written assignment covers a graph in the pupil's textbook or a reproduction of a graph drawn on the blackboard. To illustrate, the American history teacher may want to center the attention of the class on the graph, "How the American Dollar Was Spent in the War," p. 666 in Muzzey, *The History of the American People*. To check the ability of pupils to read the graph, he may ask members of the group to write the answers to the following questions:

1. Expenditures for munitions were what percentage of the total?
2. What percentage was spent for pay, food and clothing?
3. Within limits, which would the enemy prefer: to kill an American soldier or merely to disable him? Why?

The last question clearly goes beyond a direct reading of the graph. It calls for a rather penetrating interpretation of the data presented. The point involved is that if expenditures for food, pay and clothing totaled fifty-seven per cent as contrasted with only twenty-nine per cent for munitions, a country would be greatly embarrassed by having to take care of a number of wounded soldiers. Care of the wounded would necessitate approximately the ordinary expenditure for pay, food and clothing, would further necessitate the expenditure of two per cent of the total revenue for medical attention; and in addition would compel potential combatants to remain behind the lines for service in the medical corps.

It is not necessary to illustrate how similar testing procedures may be devised to measure the ability of the pupils to read a

EXAMINATIONS IN MAJOR SUBJECT FIELDS

statistical table, etc. An objective scoring of results can be assured by converting the questions asked into objective statements of the true-false or multiple-choice type.

5. Use of Books

Not only must the efficient pupil in the social studies be able to understand what he reads, but he must be able independently to use books for reference purposes. Some pupils have never learned to use the index and table of contents. The following questions are typical of many that may be asked to gain insight into the pupil's ability to use, for instance, a textbook in world history.

1. What are the major divisions into which the text is divided?
2. What is the proportionate number of pages allotted each major division?
3. How many chapters are included in the third division?
4. In what major division do you find a discussion of feudalism?
5. Give the number of the pages (inclusive) containing the account of the Protestant Revolt.
6. On what pages do you find accounts of each of the following:
 - a. The French Revolution
 - b. The revolutions of 1848
 - c. The Russian Revolution of 1917
7. Where do you find definitions of the following terms:
 - a. Rosetta Stone
 - b. Carthago delenda est
 - c. Excommunication
 - d. Industrial revolution
8. On what pages do you find evaluations of:
 - a. Bismarck as a statesman
 - b. Napoleon as a general

The ability under discussion includes much more than the ability to use effectively a single book. Among other things, it also must take into account the ability to use reference books and bibliographical aids. Space permits the inclusion of only

EXAMINATIONS IN THE SOCIAL STUDIES

a few items to suggest how to test the pupil's ability to recognize the proper reference books to be used for various purposes.

- (1) Where would one find the total amount of cotton exported annually from the United States since the World War?
 1. World Almanac
 2. Encyclopedia
 3. Reader's Guide
 4. An economics text
- (2) Where would one find a reasonably adequate discussion of the slavery controversy in the United States?
 1. An economics text
 2. An American history text
 3. An American government text
 4. Who's Who
- (4) Which would suggest a number of references dealing with the establishment of the Kingdom of Yugoslavia?
 1. A world history textbook
 2. Encyclopedia.
 3. World Almanac
 4. Reader's Guide
- (3) Where would one look in order to be able to compare the extent and boundaries of the Assyrian, Chaldean and Persian Empires?
 1. Reader's Guide
 2. Dictionary
 3. Atlas
 4. Encyclopedia

It should be noted that in exercises of this type it is both possible and desirable to include several items which are more or less correct answers. The pupil is expected to point out the best answer in each case. His doing so is one evidence that he would know the best way to spend his time in order to obtain the desired information.

6. Map Reading

This brief discussion of the testing of study skills in the field of the social studies is brought to a close with a consideration of exercises to measure the ability of pupils to interpret maps and to make accurate place locations.

Exercises may be constructed which require the pupil to think through geographical factors in a number of different settings. For example, the teacher of world history may ask the members of the class to open their books to a map of the Mediterranean area and to answer the following questions:

1. Why did Italy side with the Allies rather than with the Central powers during the World War?
2. Why did President Wilson oppose the acquisition by Italy of the entire Dalmatian coast?

It is not necessary to elaborate on how interpretative questions of this kind can be made either highly generalized or detailed and specific. In the latter instance it is possible to turn them into objective-type statements to facilitate objective grading of pupil responses.

The mastery of certain important place location facts would seem indispensable to a proper understanding of history. Pupil ability along this line may be tested for in a number of different ways. Many teachers stress free-hand sketching of maps. For example, world history pupils may be asked to sketch the Tigris-Euphrates valley, including the following detail: both rivers, the Persian Gulf, the Plain of Shinar, Babylon and Nineveh. While a certain amount of free-hand map drawing may be valuable, there is a tendency in many classes to devote to it an amount of time that is out of proportion to the values derived. The emphasis often is upon the neatness and accuracy of reproduction rather than upon the appreciations and understandings which are fundamental. Frequently the student fulfills a map assignment of this kind

EXAMINATIONS IN THE SOCIAL STUDIES

by merely tracing a map found in the textbook and without paying much attention to what the map represents.

To a considerable extent it is possible to devise objective exercises which compel the pupil to visualize and interpret a map in order to make the proper responses. The following are examples:

- | | |
|--|---------------|
| (2) An African port located west of Sicily and south of Sardinia | 1. Alexandria |
| (1) Located to the southeast of Crete | 2. Carthage |
| (4) Located directly north of Corsica | 3. Ephesus |
| | 4. Genoa |
| | 5. Marseilles |
| (2) That area of land that was located north of the Ohio River, east of the Mississippi River and south of Canada was known as the | |
| 1. Louisiana Purchase | |
| 2. Northwest Territory | |
| 3. Gadsden Purchase | |
| 4. Kansas-Nebraska Territory | |

The Northwest Territory was that area of land which was located east of the (Mississippi) River and north of the (Ohio) River.

The last two test items were included in a test given recently to 337 pupils in eighth grade American history.¹⁴ The results obtained would seem to indicate the importance of this type of diagnostic testing as well as the need for remedial instruction. Of the 337 pupils tested 219 selected the correct response in the multiple-choice exercise. On the completion-recall item, however, only 129 of the 219 were able to fill in "Mississippi" and only 106 "Ohio" in the proper blanks. This fact raises considerable doubt as to the degree of insight possessed by about half the pupils who responded correctly to the multiple-

¹⁴ Mary H. King. *Pupil Comprehension of Place Location Data in Junior High School American History*. Unpublished M.A. Thesis, University of Iowa, July, 1935, pp. 227-33, 323-27.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

choice item. It perhaps should be noted that in the total group only 150 pupils were able to write "Mississippi" in the proper blank of the completion-recall exercise. Of the remainder, 147 omitted the item; 18 wrote in "Ohio" and 22 filled in miscellaneous responses. Of the 337 pupils, only 109 wrote "Ohio" in the proper blank; 165 omitted the item; 25 wrote "Missouri"; 15 "Mississippi"; and 23 filled in miscellaneous responses. Among the miscellaneous responses to both blanks were included the following rivers: Rio Grande, Columbia, and St. Lawrence.

Perhaps the most valid measure of the pupil's ability to make place locations is a map test wherein the pupil is asked to associate a numbered or otherwise identified location with one of several place locations listed. This device has been used to a considerable extent both in standardized tests and in workbooks. It is possible for the teacher to identify place locations on a map drawn on the blackboard or on a large blackboard type map. Having done so, he may dictate a list of place locations to the pupils. The students will then respond by listing after each place location the number identifying that location on the map. This type of test was included in the study already mentioned. Of the 337 pupils tested, only 166 were able to identify the numbered location of the Northwest Territory. Of the 219 who had answered the multiple-choice exercise correctly, only 156 made the proper location. It would appear also that some were able to locate the Northwest Territory without being able to identify its natural boundary. Of the 166 pupils making the correct location on the map, only 105 were able to write in "Mississippi" and only 95 "Ohio" in the proper blanks of the completion-recall test item. Perhaps enough has been said to point out the value of different approaches in the testing of the ability of pupils to recall, recognize, and make place locations.

The Classroom Teacher and Testing

The writer senses that at this point many readers will feel that he has lost sight of reality in two directions: the amount of test construction that can be expected of a classroom instructor teaching an average of five classes and a total of perhaps 150 pupils daily; and the probable cost of mimeographing, etc., of such materials as have been described. Of these two objections the first is by far the more serious.

Perhaps the best that can be done under existing conditions is for the teacher yearly to develop rather comprehensive and adequate informal tests of study skills in connection with one subject. Once constructed, these need not be changed greatly from year to year. Suggestions for the types of exercises to be developed may be obtained from the best of standardized examinations in silent reading comprehension, etc. Furthermore, exercises such as those suggested frequently can be adapted directly from lesson plans. To the extent that the devices used permit objective scoring, it is possible to delegate this work and thus save time. In the final analysis, however, the teacher must convince himself that the need for emphasis on work skills is of such crucial importance that it cannot be neglected in either teaching or testing.

The cost of this kind of testing can be reduced to a minimum. Selections on which may be based exercises involving outlining, summarizing, etc., may be taken directly from the textbook, and therefore need not be mimeographed. The exercises themselves in some cases can be dictated or, in practically all cases, written on the blackboard. Even a map location exercise involving the use of numbered locations need not require outline maps in the hands of all pupils but may be based on a large blackboard-type outline map or even a map traced on the blackboard.

The majority of social studies teachers place comparatively

EXAMINATIONS IN MAJOR SUBJECT FIELDS

little emphasis upon the development of work skills. This situation grows out of a common misconception that pupils have mastered the skills in question when enrolled in the elementary school. The author, while he has the greatest respect for the quality of instruction at that level, does not believe that high-school or, for that matter, college students work at anything like the optimum level of efficiency. The remedy clearly involves the use of tests to diagnose pupil deficiencies, and the repeated use of the proper remedial procedures to improve pupil efficiency.

THE USE OF THE ESSAY-TYPE QUESTION

It has been demonstrated that it is possible to build objective test items which measure considerably more than the mere ability to recall isolated information. At the same time it must be admitted that the ingenuity of the objective test constructor has not been equal to the task of devising techniques for the measurement of many of the more intangible outcomes of instruction, such as to express ideas effectively in writing; to locate and organize materials independently; to pass judgment on the effect of a series of related happenings; nor has it been possible, with present objective techniques, to explore the capacity of a pupil for an unusually thorough insight into a comparatively narrow field.

What has been said, however, should not be construed as equivalent to saying that the essay examination necessarily measures these outcomes. Most of the present essay examinations in the social studies stress little else than ability to recall facts or to reproduce an organization or interpretation already provided by the textbook or teacher. Their general quality may be inferred from their characteristic approach: "Name," "identify," "describe," etc. Some essay questions which

seemingly call for an evaluation, in reality do not do so. Such a question as, "Evaluate the causes of the American Revolution," almost surely can be answered by the pupil in terms of what he can remember having read in his textbook, or having heard stated in class.

The author believes, however, that with proper care in the phrasing of questions, the essay examination can be extremely valuable in placing a premium upon values which otherwise may be neglected. The essay examination may not actually measure these outcomes, since, regardless of the quality of the question, the reliability of measurement is conditioned altogether by the accuracy of the grading of pupil answers. This is a most important point to consider in evaluating the essay-type examination as a measuring device. The claimed advantages of the essay test are often exaggerated because it seems so apparent that the questions in a good examination of this type call for the exercise of abilities that are not required from pupils in most objective tests. The point is frequently overlooked that no matter what the quality of the questions or the nature of the abilities required from the student on an essay examination, it still may prove almost worthless as a measuring device because of the failure of the teacher to evaluate pupil responses on a reliable basis.

Another advantage often claimed for the essay examination is that its use effects a saving in the teacher's time. This is largely a delusion, since whatever time is saved in the construction of the essay as compared to the objective-type test is more than offset by the greater expenditure of time necessary for the adequate evaluation of pupil responses.

However, despite the fact that the evaluation of pupil responses on the essay examination may be unreliable, it is important to stress such questions, if of the highest quality, in order that the student may be led to think and work along

EXAMINATIONS IN MAJOR SUBJECT FIELDS

desirable lines. If the essay examination is to attempt to measure such outcomes of instruction as those suggested, clearly it must be administered in a different fashion from that ordinarily followed. The pupil can neither "locate and organize materials independently" nor "express these ideas effectively in writing" in a typical testing situation. Many instructors already have taken a step in the right direction by introducing the "open-book" type of examination. Having made this concession, it would seem logical to make a further advance. The examination period is usually too short and classroom conditions too uncomfortable for effective work on a problem of major importance. The solution is the "problem-type" test which is simply the assignment of a topic for report at a future date. Thus a pupil in an economics class might be assigned the question, "Is a free-trade policy impossible for a major country today?" and told to submit his findings a week or two later.

A limitation of the objective examination is that it demands that all pupils demonstrate their reasoned understanding of subject matter in terms of a common body of elements. This provides no opportunity for measuring the pupil's depth of insight into any single phase of history. Superior students often do a surprising amount of reading on a topic which especially interests them. The writer recently had in his tenth grade class in world history an exceptionally able boy whose father had been an officer in the World War. When the class began the study of the topic, "The Background of the World War," this pupil's interest was especially aroused by the question of "war guilt." With little or no urging on the part of his teacher, he read the authoritative works by Fay, Barnes, and Schmitt on this subject, and demonstrated a surprising grasp of the problem as a whole. Such depth of understanding, and the warmth of interest which it connotes, is certainly a

EXAMINATIONS IN THE SOCIAL STUDIES

desirable outcome of instruction in the social studies. The old-fashioned term paper, where the topic is reasonably restricted and the materials available, offers one approach to the testing of such an outcome.

The essay question also may be so phrased as to compel pupils to think through a series of related events, and to express their conclusions in terms of what they have read. Thus the following question, "How did the Glorious Revolution of 1688 affect the political destiny of Louis XIV?" almost certainly cannot be answered in terms of a direct quotation from the pupil's textbook. Most likely the pupil would have to review how Louis, in attempting to gain the Rhine frontier, had already clashed with the Stadtholder of Holland. The Revolution of 1688 ousted the Stuarts, pensionaries of the French king, from the English throne. With the added prestige and power coming from his accession to the English throne, William of Orange, already the implacable foe of Louis, could proceed more effectively than ever to arrange coalitions to arrest French expansion. A question such as this certainly would reveal great differences, both in grasp of fact and in interpretation, from pupil to pupil.

Many studies have reported at length the unreliability of grading when essay questions are used. These studies nearly always describe a situation where the reader of the test was trying to assign questions some numerical value with a possible score of 100 for a "perfect" paper. The weakness of this system of grading is obvious. It is impossible, for example, to distinguish eleven separate categories ranging in value from 10 to 0 in evaluating pupil responses to an essay question. The teacher may save time and avoid claiming fictitious reliability for his test by limiting the categories to four: 3 for a superior answer, 2 for an average, 1 for an inferior, and 0 for no answer at all. Even this system of grading has a major weakness

EXAMINATIONS IN MAJOR SUBJECT FIELDS

when applied to examinations consisting of several questions. The teacher grading the examination may begin his work with an exaggerated notion of what may be expected, and assign low values to the questions on the first few papers. Realizing that he is being too severe, he gradually begins to raise his grades. In consequence, the authors of the first papers graded have been unfairly penalized.

If the examination consists of several questions, it would seem best to grade the first question on each paper before evaluating the second question on any paper. In this way the reader can make a direct comparison of the answers returned by several pupils to the same question and will find it relatively easy to classify these as superior, average, and inferior. As he reads the answer to the first question, he can place each paper in one of four piles of papers in front of him, i.e., superior, average, inferior, no answer. In picking up the examination papers to read the answers to the second question, he will of course read the papers in different sequence. He thereby avoids any systematic variations in grades assigned resulting from a tendency to be unreasonably "hard" or "easy" in the scoring of the first few or last few papers. In view of the fact that any teacher is almost certain, consciously or unconsciously, to be prejudiced in favor of or against certain pupils, it may facilitate objective evaluation of pupil work to have the identification made by code number rather than name.

After the pupil responses to the several questions on the examination have been evaluated and assigned a numerical grade on the scale 3, 2, 1, and 0, the next step to be taken by the teacher depends on what he conceives to be the function of the test. If he considers it a diagnostic examination, he must rule that all answers assigned "inferior" ratings, and conceivably even those assigned "average" ratings, are unsatisfactory. Each of these would seem to indicate a need for

EXAMINATIONS IN THE SOCIAL STUDIES

reteaching and retesting. On the other hand, if the essay examination is intended as a general achievement test, the pupils can be ranked in order of achievement by totaling the numerical values assigned each question. Thus, on an examination consisting of five questions each graded in terms of a four-point scale (3, 2, 1, 0), the highest possible score would be 15; the lowest 0. The teacher can conveniently make a frequency distribution of the numerical grades and assign letter grades (A, B, C, D, Fd.) if he sees fit.¹⁵

In conclusion, the following suggestions may be made relative to the use of the essay-type examination:

1. Because of its greater reliability and more extensive sampling per unit of testing time, use the objective-type examination to measure all abilities and skills which can be so tested.

2. Make use of essay-type questions of highest quality in order to place emphasis upon outcomes of instruction which are commonly neglected in objective testing.

3. To stress the development of desirable skills, use the essay type of exercise as a problem-type test or research paper rather than as an instrument in a typical testing situation.

THE TESTING OF ATTITUDES

Recurring emphasis in certain professional literature on the supreme importance of the development of "desirable emotion-alized attitudes, ideals, and modes of behavior" through instruction in the social studies has created a demand for the "measurement" of such ultimate objectives. Usually the modes of behavior mentioned are those suggested by such general terms as "open-mindedness," "sympathy," "cooperation," "honesty," "tolerance," etc.

¹⁵ A method of changing numerical grades into letter grades is suggested on pp. 118-25.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

It seems unlikely that it is either possible or desirable to establish a set pattern of behavior based on generalized attitudes of this type, and intended to operate in all of the wide variety of specific situations a child will meet. For instance, one may wish a boy to be a team-worker on the football team, but not in the marauding expeditions of a neighborhood gang. It must be recognized at the outset that the so-called "desirable attitudes" consist in reality of specific attitudes and specific ideals with reference to specific situations or types of situation.

It should be equally apparent that it is not the sole responsibility of the social studies to develop all of the specific attitudes, ideals, and modes of behavior implied in such general terms as "honesty," "cooperation," "sympathy," etc. In a single day the pupil may develop a number of specific modes of behavior suggested by the general term "cooperation" in both in-school and out-of-school situations. He may eat his breakfast on schedule; he may come to school on time; he may be prepared to participate in group discussion in the history class; he may do some expert "blocking" on the football field, etc. Clearly it cannot be the sole responsibility of the social studies teacher to measure the development of specific attitudes and modes of behavior developed in such a variety of situations.

At the same time it must be granted that there are a number of social institutions, problems, and practices the study of which properly pertains to social science and toward many of which it is quite possible to develop emotionalized attitudes on the part of pupils. But even so, there can be no question of measurement unless some supreme authority decides irrevocably what is the proper attitude to be developed with respect to a given problem. For example, what is the generalization with respect to the ownership of public utilities which the pupil should be taught to accept and concerning which his convictions should be emotionally intensified? Manifestly,

EXAMINATIONS IN THE SOCIAL STUDIES

the very term "social problem" suggests that the issue is controversial. Furthermore, changing conditions may modify even what was once quite universally accepted as correct and proper. Contrast, for example, the typically American attitude toward social insurance prior to the disappearance of the frontier and today.

There may even be a danger in the development of uniformity in attitudes, ideals, and modes of behavior, since the ultimate result would be a dead level of conformity. The inertia of stability can lead to stagnation and render progress difficult. Perhaps the real function of social studies instruction is to offer the pupil a rationalized basis for formulating tentative conclusions. He should master basic concepts and generalizations pertinent to the solution of current problems, but his general attitude, if any, should be one of reserving final judgment; of standing ready at any time to accept and to integrate new information and new ideas which may require revision of his tentative conclusions.

Finally, it should be pointed out that even if it were possible to describe in terms of a position along an arbitrary scale a pupil's attitude toward, for instance, private ownership of public utilities, and even if it were possible to designate one end of the scale as that on which the pupil's answer should be located, it still would not follow that his attitude in any real sense had been "measured." As soon as the pupil senses that his responses are being evaluated in terms of correct and incorrect, his chief concern is to make the "correct" response. Whether or not he really has accepted the opinion he has expressed, it is impossible to tell.

This chapter must end on the same note that it began on. It has been written to supplement the chapters on the theory and practice of test construction. The writer feels that the

purpose of all three chapters is to provide teachers with a basis for evaluating published examinations and with specific suggestions for improving tests of their own construction. Though it is hoped that the classroom teacher will never weary in refining his informal tests, it is perhaps equally important that he not lose sight of the fact that the written examination as here discussed is only one of many devices available for the appraisal of instruction.

QUESTIONS FOR DISCUSSION

1. What are the purposes for which a general achievement type of test may be used?
2. What sources contain authoritative descriptions of content on which a general achievement test may be based? What are the restrictions on the elements of content which may be included in this type of examination?
3. What is a table of specification? Why must several independent bases for the classification of test items be used in order to insure a good sampling?
4. Why is it desirable to emphasize the testing of the pupils' reasoned understanding of significant information, relationships, and generalizations? How does this emphasis differ from that which often characterizes both objective and essay-type examinations?
5. What is meant by "homogeneous grouping" in matching exercises? Why is this important? How may it be insured?
6. Why is it important to avoid "pat" statements and textbook phraseology in the construction of test items? What is meant by irrelevant "clues"? Why should care be taken to eliminate these from test exercises?
7. How does the function of the diagnostic test differ from that of the general achievement examination? How does this difference affect the theory of sampling which may be followed in building diagnostic tests? How shall the teacher interpret pupil scores on this last type of test?

EXAMINATIONS IN THE SOCIAL STUDIES

8. Why should a teacher include exercises dealing with work skills in his testing program? What use should be made of pupil scores on such tests? What are some of the important skills necessary to effective learning in the social studies?
9. What outcomes of instruction is it difficult or impossible to measure by objective-type tests? To what extent can these be measured through the use of essay questions? What are the advantages of "open-book" and "problem-type" tests?
10. How may the grading of essay questions be facilitated?
11. What are some of the difficulties in the testing of attitudes?
12. Prepare a table of specifications for a general achievement examination in one of the social studies.
13. What types of exercises are best adapted for use in social studies testing? Explain why, and in which specific situation, each is most effective. Build an example of each type of exercise discussed.
14. Find an example from published tests of a matching exercise which is not homogeneous. Explain *why* you believe this to be the case. Revise the exercise to make it homogeneous.
15. Build a homogeneous matching exercise. Explain in detail why, in your estimation, it is homogeneous.
16. Find examples of published test items which contain irrelevant clues. Revise the items so as to eliminate these. Do similarly with items characterized by "pat" or textbook phraseology.
17. Build several multiple-choice items which you believe are free from structural defects. Explain why you believe this is the case.
18. Build a diagnostic-type test to measure pupil mastery of one of the important work skills in social studies.
19. Prepare five essay-type questions which you believe measure outcomes of instruction that cannot be tested by means of objective tests. Explain in detail why you believe that your questions conform to this standard.

CHAPTER V

EXAMINATIONS IN THE NATURAL SCIENCES

TYPES OF OBJECTIVES IN THE NATURAL SCIENCES

TEACHERS of the natural sciences are concerned with bringing about a variety of changes in the behavior of students. The desired outcomes of a course in the natural sciences probably include: the acquisition of a knowledge of the principles and facts of the course; an understanding of the important technical terminology and symbols used in this field; the ability to identify structures and processes and their functions, as, for example, in botany and zoology; a familiarity with reliable sources of information on science problems; the ability to recognize unsolved problems in science; an interest in natural phenomena and an interest in solving problems in natural science for which the student has no present solution; the ability to draw reasonable generalizations from experimental data; the ability to plan experiments to test hypotheses; the ability to apply scientific principles to situations new to the students; skill in laboratory techniques. In addition to the objectives which may fall almost wholly within the field of the natural sciences, many teachers are also concerned with certain objectives which are partly the function of science courses and partly the function of other courses. These include such outcomes as the ability to prepare effective reports both orally and in writing, the habit of carrying science attitudes and abilities into the social studies and other non-science courses, an attitude of tolerance toward new ideas, and the habit of cooperation with others.

EXAMINATIONS IN THE NATURAL SCIENCES

The purpose of a course in the natural sciences is to bring about changes in the behavior of students in the directions of the objectives appropriate for this course.¹ Any satisfactory examination program must obtain evidence of the degree to which these changes are taking place. What constitutes evidence as to the degree to which these outcomes are resulting? How can this evidence be effectively collected? These are the major problems in preparing examinations for the natural sciences.

EXAMINATION PROCEDURES APPROPRIATE FOR EACH OF THESE TYPES OF OBJECTIVES

1. Acquisition of Information

The acquisition of information, apart from its use, is not an important end in itself. Although the ability to repeat facts which have been remembered is sometimes demanded in life, the greater significance of information is its use in extending meanings and in solving problems. In order to analyze the achievement of pupils, or in order to diagnose particular strengths or weaknesses, teachers often wish to test acquisition of information separately from the pupil's facility in the application of information.

To determine appropriate methods for getting evidence of the degree to which pupils have acquired important information, it is necessary to define this objective in terms of the behavior of pupils who have "acquired important information." If the acquisition of information means to be able to state the desired facts when direct questions are asked, then the way to test this behavior is to ask these direct questions and expect the student to state the information desired.

¹ In this chapter the terms "objective," "desired outcomes," and "desired changes in behavior" are used synonymously. The term "behavior" as used here means any sort of appropriate reactions of students — mental, physical, emotional, and the like.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

If the acquisition of information is defined as the ability to distinguish between facts and misconceptions in the material which the pupil reads and in the lectures and discussions which he hears, then the way to test this behavior is to present him with appropriate written and oral material and ask him to indicate which statements are facts and which are misconceptions. This latter definition requires the further decision in each case as to the desired degree of fineness in discrimination between fact and misconception.

A still more significant definition of the acquisition of information is the ability to use the information in enlarging one's conceptions. This means that the pupil goes beyond memoriter reproduction of facts. He is able to indicate the significance of those facts. This definition implies that the test of behavior will involve requiring the pupil to state the information in his own words and to indicate its more significant implications. In all of these definitions the teacher will need to determine the kinds of stimuli which he believes would bring forth the desired response from the pupil. In some cases these stimuli would be specific questions, and in others more general problems.

In a test of the pupil's ability to state desired information in response to direct questions, it is important that the questions be couched in phraseology which is not a verbatim reproduction of that of textbook or classroom. Each question should involve information the recall of which is significant, and should be phrased in the way in which the direct questions might normally arise in life. Thus, the question, "What is the statement of Boyle's Law?" is not very effective because it uses phraseology almost identical with that of the textbook and does not give the pupil a chance to show his ability to recall information when any reorganization of the facts is involved. The questions. "How is the volume of a kilogram of

EXAMINATIONS IN THE NATURAL SCIENCES

air affected by changes in the pressure applied to it? Upon what other gases do changes in pressure produce similar effects?" are better means of testing this objective because the pupil is expected to show some understanding of the meaning of the words of Boyle's Law and because the situation is more like those in life in which the pupils may be expected to recall information in response to direct questions.

Exercises of this sort are sometimes called "essay questions" although the answers to such information questions obviously involve a minimum of composition. Essay tests have been criticized because different readers of the pupil's answers often give very different evaluations to the responses. Essay tests have also been criticized because the sample of behavior tested in a class period is frequently too small to be a reliable indication of the pupil's achievement. Both of these weaknesses can more easily be minimized in acquisition-of-information tests than in tests for some other type of behavior.

Much of the variation among individual readers in the evaluation of the same behavior is due to a failure to agree upon the objective being tested and to clarify the meaning of this objective in terms of the desired behavior. Considerable variability is frequently due also to lack of a common scale of words or numbers to express the judgments of the different readers. In the case of the information given in response to direct questions, specifications can be prepared to guide the reader in his evaluation. The specifications indicate the factors to be considered in judging the response to this kind of exercise and also the scale of values to be used. When the reader follows such a set of specifications, his evaluations are quite objective and will not usually fluctuate markedly from those of another competent reader who follows similar specifications. A portion of a set of specifications follows.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

Specifications for Evaluating Facility in Stating Information in Response to Direct Questions

Purpose of the test

This part of the test is to obtain evidence of the student's acquisition of information. If you are able to identify the meaning of his statements, do not consider his spelling, handwriting, grammatical construction, or method of presentation in your evaluation.² Judge only the accuracy and adequacy of the information he stated.

Values to be assigned to answers

Question 1a — How is the volume of a kilogram of air affected by changes in the pressure applied to it?

Allow 4 points credit for answers which indicate direction and nature of relation and constancy of temperature; for example, "Volume varies inversely as pressure if temperature remains constant."

Allow 3 points for answers which indicate direction and nature of relation but do not mention constancy of temperature; for example, " $\frac{V_1}{V_2} = \frac{P_2}{P_1}$ " or an equivalent expression.

Allow 2 points for answers which indicate direction of effect but not amount; for example, "Volume decreases as pressure increases."

Allow 1 point for omissions, or "I don't know" statements. This assumes that the teacher considers it better for a student to recognize that he does not know than to be mistaken in the facts he thinks he knows.

Allow no credit for mistaken conceptions, such as "Pressure has no effect on volume," "Increasing pressure increases volume."

² The purpose of this instruction is to prevent the judgment of ability to state information from being influenced by the judgment of ability to write effectively. When the latter is also to be considered, the objectivity and the diagnostic value of both evaluations may usually be improved by making the two judgments independently.

EXAMINATIONS IN THE NATURAL SCIENCES

Question 1b — Upon what other gases do changes in pressure produce similar effects?

Allow 2 points credit for answers which indicate that all or nearly all gases are affected in similar fashion.

Allow 1 point credit for omissions, or "I don't know" answers.

Allow no credit for mistaken conceptions, such as, "No other gases," "Gases which are lighter than air," "Gases where molecules are composed of only one atom."

This set of specifications for evaluating the student's responses is only suggestive of the type which each teacher will want to make for himself. The same result would be obtained if no credit was allowed for omissions and points were deducted for mistakes. The number of credits assigned is not significant, but it is important to make the scale for judging each answer as definite as possible so that all students with the same or equivalent answers will receive the same scores.

The problem of finding time for a reliable sample of questions for such an information test is not serious because questions of this sort take little time to answer and require little writing. A fairly large number may be included in the average examination period. It should be clear, however, that this statement refers only to the information test. Since there are also other objectives to be tested, this type of examination represents only a portion of the complete examination program in science.

In a test of the pupil's ability to distinguish between facts and misconceptions in material which he reads and in the lectures and discussions which he hears, it is possible to use written and oral material, both in context and in separate statements. Articles and items from daily and Sunday newspapers and from "popularized science" periodicals are convenient sources of written material. When these are to be tested in context, a complete article or an adequate portion of it can be

given to the student. He is asked to read the article and to indicate those statements which he believes to be untrue and also those about the truth of which he is uncertain. When statements without context are to be included in a test, they may be reproduced in the form of a true-false test. The student is asked to mark each statement to show whether he believes it to be true or false or whether he is uncertain. Local conversations and discussions among pupils and adults provide suggestive statements for oral presentation to the pupils.

In evaluating such test exercises, the following portion of a set of specifications is suggestive.

*Specifications for Evaluating Ability to Discriminate between
Facts and Misconceptions*

Purpose of the test

This part of the test is designed to obtain evidence of the student's accuracy of information by determining his ability to discriminate between facts and misconceptions.

Values to be assigned to answers

Allow 3 points for every fact which the pupil correctly identifies as true, for every misconception which he correctly identifies as such, and for every doubtful statement which he correctly marks as uncertain.

Allow 2 points for every fact which the pupil marks as uncertain, for every misconception which the pupil marks as uncertain, and for every omission.

Allow 1 point for every uncertain item which the pupil marks as a fact or as a misconception.

Allow no credit for any fact which the pupil marks as a misconception, or for any misconception which the pupil marks as a fact.

In a test of the pupil's ability to state information in his own words and to indicate the more significant implications of facts, it is important to use facts which do have significant

implications. This definition of acquisition of information differs from the definition of ability to interpret new data only in the use of data with which the pupil is already familiar. It is also closely related to the application of generalizations in that it involves the relation of facts as well as their formulation. The examination questions should be couched in phraseology different from that of the textbook or classroom and should ask for implications which will show that the pupil understands the meanings of the facts — implications which have not been specifically suggested to him before. For example, the question, "What is the 'nitrogen cycle,' and what is its significance?" is couched in textbook phraseology and requires only memoriter reproduction of textbook statements. The questions, "From what sources do green plants get nitrogen? From what sources do animals get nitrogen? Are the supplies of available nitrogen for green plants and animals decreasing? Why?" are better for most biology students, since the answers require more reorganization of facts and less memoriter reproduction.

In evaluating this type of test, two methods are often used. One method does not require the preparation of a key in advance of grading the papers. The reader evaluates the first exercise in all the papers, then the second exercise in all the papers, and so on, instead of attempting to grade one pupil's paper all the way through, then another pupil's entire paper, and so on. The following general directions illustrate the procedure in grading an exercise.

*General Directions for Evaluating Recall of Facts and Statements
of their Implications*

Purpose of the test

The purpose of this test is to determine the pupil's acquisition of information and his understanding of its meaning. If you

EXAMINATIONS IN MAJOR SUBJECT FIELDS

are able to recognize the meaning of what the pupil has written, do not consider his spelling, handwriting, grammatical construction, or method of presentation in your evaluation. Judge only the accuracy and adequacy of the facts and implications which he has written.

Values to be assigned to answers

Judge all the answers to the first question, then all the answers to the second question, and so on. Use five degrees of quality in indicating your evaluation. Allow 4 points credit for the best answers, 3 points for those in the next degree of quality, 2 points for those of average quality, 1 point for those in the degree just below average and no credit for those of poorest quality. When you are uncertain as to the degree of quality of an answer, lay it aside until you have evaluated the others, then reread it and judge it as accurately as you can.

When a large number of papers are to be evaluated, or when the accuracy of the grading is particularly important, it is better to prepare a grading key before scoring the papers. Upon the grading key will be indicated the answers expected for each question and the numerical values to be assigned. The following portion of a grading key illustrates its nature.

Grading Key for Test of Recall of Facts and Statements of their Implications

Question 1 — What are the effects of lowering the resistance of a circuit conducting electricity at constant voltage? How are these effects used in protecting the household from damage by electricity?

Allow 4 points for such accurate and complete answers as the following: "Increases the intensity of current (or amperage), increases the amount of electrical energy transformed into other kinds of energy, increases the heat of the circuit. This last fact is used in the construction of fuses which melt when the intensity of current becomes excessive."

EXAMINATIONS IN THE NATURAL SCIENCES

Allow 3 points for answers which mention only the increase in heat and its significance in fuses.

Allow 2 points for answers which mention one or more effects but do not state the use in fuses.

Allow 1 point for omissions.

Allow no credit for answers which contain significant errors of fact.

A promising method of collecting evidence about information acquired is a multiple-response or selection-of-facts exercise in which the student is to place a + mark opposite each statement which is true and to avoid marking the misconceptions. For example:

Some characteristic properties of digestive enzymes are:

- a. They hasten chemical changes which occur in food. (+) a.
- b. Variations in alkalinity or acidity have no effect upon the rate of enzymic action. () b.
- c. Enzymes are rendered inactive at extremely high or extremely low temperatures. (+) c.
- d. Any one enzyme acts only upon certain specific types of foods. (+) d.
- e. Any one enzyme is capable of acting upon any type of food. () e.
- f. They are changed during a chemical reaction and remain changed at the end of the reaction. () f.

In this kind of exercise the student may check as many ideas as he thinks are true. Statements a, c, and d are statements of scientific principles which have been derived from experimental evidence. Statements b, e, and f are misconceptions which some students entertain about the properties of enzymes. The statements in the list should afford students an opportunity to indicate what they know about the properties of enzymes.

Another possible kind of exercise is as follows:

- Digestive enzymes: () 1. Are changed during a chemical reaction and remain changed at the end of the reaction;
 () 2. Are not affected by variations in temperature;
 (+) 3. Hasten chemical changes which occur in food;
 () 4. Are not affected by variations in alkalinity;
 () 5. Are not affected by variations in acidity.

In this kind of exercise the student is asked to consider the five ideas and to check only the one which he thinks is true. In both of these exercises prevalent misconceptions should be included so that the students will have an opportunity to make careful discrimination between misconceptions and established facts. When the exercises are so constructed, an analysis of the results will help the teacher to discover the misconceptions of the students in the class and afford a basis for directing the study of individual pupils and for adapting teaching materials and procedures. These exercises can be administered and scored more quickly than the essay tests. When the teacher can find on the market such tests which cover the important information in his course, the considerable time required to build these tests is saved.

2. Understanding of Important Technical Terminology and Symbols

A second objective commonly stated for courses in the natural sciences is the understanding of the technical terminology and symbolism. Although it is a phase of the acquisition of information, its significance entitles it to separate treatment. The best type of examination for testing the degree of attainment of this outcome can be determined only when the objective is defined in terms of the behavior desired.

The kinds of behavior which are called "understanding of terms and symbols" are somewhat varied. Certain terms

EXAMINATIONS IN THE NATURAL SCIENCES

students may be expected to use correctly in their own written and oral reports. The test of this behavior is to make a periodic check of the written and oral reports of students, noting whether these terms are used, whether they are used with the proper meaning, whether they are spelled correctly in the written reports, or pronounced properly in the oral reports. Such a test requires a list of terms which pupils may be expected to use in their own reports.

Pupils may be expected to define certain terms in their own words. In such cases, the direct test is to present the terms and ask the pupil to define each one in his own words. The major problem in evaluating this type of test is to determine what are acceptable definitions for the terms. For some terms, rather general descriptions may be adequate. In some cases only the general area in which the term is classified may need to be identified. When a key has been made out indicating the kind of definition which is acceptable for each term, the evaluation is somewhat simplified. Many teachers believe that it is worse for a pupil to hold an erroneous idea of a term than to recognize that he does not know its meaning. If this is true, the omitted items should receive a higher value than those for which the pupil gives an erroneous definition.

The student may also be tested on his ability to recall a term when confronted with its definition or description. This test is easily administered. The pupil is given an appropriate series of definitions or descriptions and is asked to write after each one the term which it most accurately defines or describes.

There are some additional terms which the pupil may need to recognize when he encounters them in his reading, although he may not need to remember how to spell them or be able to define them out of context. This kind of behavior is perhaps best measured by a reading test covering typical science

EXAMINATIONS IN MAJOR SUBJECT FIELDS

material. Questions which demand an understanding of these terms in the reading context require some skill in formulation.

Examples of some exercises which give evidence of the understanding of technical terms are as follows:

Directions: Below is a numbered list of botanical terms arranged in alphabetical order. Following the list are some definitions or descriptions of terms used in botany. Read each definition or description, decide what term it is, then place the number of the term in the parentheses after the definition or description.

- | | |
|------------------|--------------------|
| 1. Anthocyan | 16. Osmosis |
| 2. Carbohydrates | 17. Parasite |
| 3. Chlorophyl | 18. Photosynthesis |
| 4. Cilia | 19. Plasmolysis |
| 5. Dessication | 20. Pollination |
| 6. Digestion | 21. Proteins |
| 7. Fermentation | 22. Respiration |
| 8. Fertilization | 23. Saprophyte |
| 9. Germination | 24. Scion |
| 10. Guard cells | 25. Spore |
| 11. Host | 26. Stock |
| 12. Hydrophyte | 27. Transpiration |
| 13. Imbibition | 28. Turgidity |
| 14. Mitosis | 29. Variations |
| 15. Mutations | 30. Xerophyte |
- a. The process by which water is lost from the cells of a plant by evaporation into the intercellular spaces, and the escape of the vapor through the stomata. (27) a.
- b. The collapse and contraction of the protoplasm of a cell due to the passing of water out of the cell. (19) b.
- c. An organism upon or within which other organisms live and receive food. (11) c.

The provision of more words in the list than there are definitions and descriptions helps to eliminate successful guessing

EXAMINATIONS IN THE NATURAL SCIENCES

in answering the questions. The examination exercise can be modified by choosing the words to be included in the list. For example, the distinction between "pollination" and "digestion" is marked, while the distinction between "pollination" and "fertilization" may not be so clear to students although very apparent to botanists. Hence, fine distinctions can be made by choosing appropriate words for the list.

In the next exercise the term which best represents the definition or description is to be checked.

The process by which water is lost from the cells of a plant by evaporation into the intercellular spaces, and the escape of the vapor through the stomata is known as: () 1. Respiration
() 2. Osmosis () 3. Imbibition (+) 4. Transpiration
() 5. Plasmolysis.

In the exercise below, the statement which best defines or describes the term is to be checked.

Transpiration is:

- a. The process by which two miscible liquids or gases or solids in solution become equally mixed throughout..... () a.
- b. The process by which water is soaked up by a plant..... () b.
- c. The collapse and contraction of the protoplasm of a cell due to the passing of water out of the cell..... () c.
- d. The process by which liquids and substances in solution tend to pass through permeable and semi-permeable membranes () d.
- e. The process by which water is lost from the cells of a plant by evaporation into the intercellular spaces, and the escape of the vapor through the stomata..... (+) e.

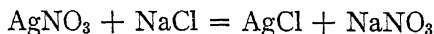
In chemistry courses, students are expected to associate certain symbols with the chemical elements and compounds

which they represent. This ability may be tested by asking the students to give the symbols for an element or compound, or by presenting the symbols and asking them to name the element or compound. Paper-and-pencil exercises like the following may be constructed.

Directions: Below is a list of elements, compounds, symbols and formulae. After each, write the symbol, formula, or name of the element or compound represented.

Iron.	_____
Sulfuric acid.	_____
Mg.	_____
NaNO ₃	_____

Further measures can be obtained by asking them to write the chemical equation in symbols for the reaction between, for example, zinc and sulfuric acid, without consulting any sources of information. Again, they may be asked to read an equation such as the following by naming the compounds.



Their behavior in these situations gives some evidence of their knowledge of the names and symbols of elements, and of the names and formulae of compounds as used in equations.

3. Ability to Identify Forms, Structures and Processes and to State Their Functions

A third kind of informational objective is the ability to identify forms, structures, and processes and to state their functions. In biology courses, students may be expected to identify plant and animal forms. They may be expected to be familiar with plant and animal structures and processes, and with the functions of these processes. In physics courses, students may be expected to be familiar with various machines, and to identify their parts and the functions of those parts. In chemistry courses, students may be expected to be ac-

EXAMINATIONS IN THE NATURAL SCIENCES

quainted with various chemical processes, with the different structures involved in those processes, and with the functions of the structures. In botany courses, students may be expected to name and identify the plants and flowers which they may observe in the fields and woods. A possible paper-and-pencil examination of this ability consists of photographic reproductions of local plants, which the students will identify by writing the name of each plant.

Students may be expected to state the names of structures of actual animals. For example, each student may be shown a frog dissected so that the heart is exposed and may be asked to name each structure of the heart and its function as it is pointed out. Or a numbered string may be attached to each of the various heart structures, and the students asked to state the name and function of each structure thus designated. A paper-and-pencil examination may consist of photographic reproductions or life-like drawings of animals, with their important structures indicated by a numbered guide line; at the side of the reproduction may be placed a list of structures to be identified. The students would respond by placing the diagram number after the name of each structure to be identified. A list of functions of the various structures may be provided and the students directed to place after each function the number of each structure which performs it. A sample exercise is shown on page 230.

Similarly, in physics, diagrams or pictures of machines or other physical devices may be used. The parts of these devices are numbered, and the student is directed to place the appropriate number after the name of the part. The functions of these parts can also be listed and the students asked to place after the function the numbers of the parts which assist in its performance. In chemistry, the processes of manufacturing can be depicted in diagrams or pictures and the students given

EXAMINATIONS IN MAJOR SUBJECT FIELDS

opportunity in similar fashion to identify the stages in the process and the functions of those stages.

Heart of Frog (Ventral View)

Structures:

Heart of Frog (Ventral View)	a. Carotid artery.....	5 or 6	a.
	b. Left auricle.....	2	b.
	c. Post caval vein.....	4	c.
	d. Pulmocutaneous artery.....	8	d.
	e. Right auricle.....	10	e.
	f. Superior vena cava.....	9	f.
	g. Systemic artery.....	7	g.
	h. Truncus arteriosus.....	11	h.
	i. Ventricle.....	3	i.

Functions: structures

which function in

j. Pumping and conveying
blood away from the
heart.....

3, 5, 6, 7, 8, 11 j.

k. Receiving blood from
parts of the body
other than the lungs.

4, 9, 10 k.

l. Receiving blood from
the lungs.....

1, 2 l.

The ability to identify the stages, structures, and functions of processes such as mitosis, embryological development, life histories, and so on, is another objective of some courses in biology. Consider the process of mitosis. When students actually observe stages of mitotic cell division under the microscope, or in a microphotographic movie of the process, they are expected to name the stages, structures in these stages, and functions of the stages and structures. Again, a promising exercise consists of numbered photographic reproductions or lifelike drawings, treated as before. Care must be taken that

in each reproduction the structures are as lifelike as possible. Pictures or diagrams taken from the student's textbook should not be used, since they would require but a low level of recognition. The element of guessing may be controlled by numbering more structures in the diagrams than are given in the list of structures. Thus, when the student is ready to identify the last structure in the list, he cannot conclude that the answer is the only other available structure not identified up to that point.

4. Familiarity with Reliable Sources of Information on Science Problems

In every science field, the published information is voluminous, and additions are continuously being made. Students could not possibly be expected to remember all of it that is useful to them. They must know, however, where to find new facts as these become known. Unfortunately, there are available many sources which are not dependable. Advertisers, sensational newspapers, radio, motion pictures, popular magazines, publications purporting to be scientific but containing little more than pseudo-science are more widely known than the more reliable sources. Familiarity with dependable sources of information is therefore an important educational objective.

There are two major aspects of behavior involved in this objective. The pupil may be expected to make some investigations which will require the collection of information from dependable sources. The sources that he consults may be observed as he works in the laboratory and the library. In his reports of his projects, both the references cited and the information selected throw light on the degree of his familiarity with sources.

The second major aspect of the pupil's behavior concerns

EXAMINATIONS IN MAJOR SUBJECT FIELDS

the habit of utilizing reliable sources of information and of discounting information suggested by unreliable sources as he meets with problems of daily life which involve scientific knowledge and as he extends his own ideas and concepts of the natural world. This aspect emphasizes the use outside the school of sources of science information. Direct evidence of the development of this habit may be obtained by analyzing the pupil's weekly record of free reading done, by noting his conversations with other pupils, and by interviewing him personally.

Some of these methods of testing are so time-consuming or require such an intimate acquaintance with the pupil as to be impracticable with larger groups. There are three types of examinations which have been found in some classes to give results very similar to those obtained from direct observations. The first type is an "essay test." The pupil is asked to write the sources which he thinks are likely to be the best to use in seeking specified information or in investigating certain problems. The following sample is illustrative.

Directions: In each of the exercises below, you are to suggest the sources which you think are best for getting information on the question given. Be as definite as you can in your suggestions. If you mention a book, or magazine, or newspaper, state its name. If you do not know its title, tell how you would find it. If you suggest some other kinds of sources, be just as definite in describing them.

1. Where could you find out about the general principles which help to explain the methods of sending pictures by wire?
2. Where could you determine the relative electrical conductivity of iron, copper and aluminum?
3. If you were making a special report on the corpuscular theory of light, where would you get helpful information?

The questions may involve problems which the pupil might encounter in either his school work or his out-of-school life.

EXAMINATIONS IN THE NATURAL SCIENCES

In evaluating the answers, the following sample set of directions may be helpful.

General Directions for Judging Pupil's Statements of Reliable Sources of Information

Purpose of the test

This test is used to obtain evidence of the pupil's knowledge of dependable sources for various kinds of science information. Judge only the value of the sources he suggests. Do not attempt to evaluate the quality of his language or the neatness of his paper.

Values to be assigned to answers

Allow 4 points credit for each source listed by the pupil which is reliable for the information sought and which is as available as any other equally reliable source.

Allow 3 points credit for each source listed by the pupil which is reliable for the information sought but is not as available as other reliable sources that he failed to mention.

Allow 2 points credit for each source listed by the pupil which is likely to contain some of the information sought but is only fairly dependable.

Allow 1 point credit for each source which is so vaguely defined by the pupil that he would be unlikely to find it without considerable loss of time. Also, allow 1 point for an omission or an "I don't know" answer.

Allow no credit for any source listed by the pupil which is unlikely to provide any helpful information or which provides information that is not dependable.

The second method is to give the students a numbered list of sources of scientific information, and direct them to indicate by number those which are best to use in seeking specified information or in investigating certain problems. This method differs from the one just described only in that the pupil selects the sources from a given list instead of recalling them independently. If the test is to be effective, the list of sources

EXAMINATIONS IN MAJOR SUBJECT FIELDS

should include all those which pupils are likely to use, both reliable and unreliable. In preparing a scoring key, the possible answers are evaluated as in the method just described; i.e., the amount of credit given for each answer may vary from 4 to 0 depending upon the reliability and availability of the source suggested.

The third test consists of a similar series of questions calling for specified information. Below each question are listed several sources, usually from 4 to 10. The student is asked to rate every source in relation to the desired information, marking with a 1 those which he believes to be good, with a 2 those which he believes might possibly be helpful, and with a 3 those which he considers not worth consulting. The sources listed should include, not only those which are reliable, but also those which are likely to be used even though they are not reliable or helpful. The scoring key for this type of test should contain the teacher's rating of each source.

A possible method of scoring, based on the assumption that responses agreeing with the key should receive the highest score while responses furthest from the key should receive the lowest score, is as follows:

		<i>Key</i>		
		1	2	3
Student's Paper	1	+ 2	- 1	- 2
	2	+ 1	+ 2	- 1
	3	- 2	- 1	+ 2

This diagram means that if a source of scientific information on the student's paper is marked 1 and the key indicates 1, the score received is + 2. If the source on the student's paper is marked 1 and the key indicates 3, the score received is - 2; two points are subtracted. Omitted responses are scored zero. A similar method may be used to evaluate the responses in the other kinds of exercises. Other values may be inserted in the squares if they are found to give a higher correlation with the kind of behavior which is accepted as expressing the objective.

5. Ability to Recognize Unsolved Problems in Science

Another objective of science teaching involves the student's ability to recognize the large or somewhat vague problems in various life situations for which the student does not, at the time, have a solution, and also his ability to define the more specific questions which need to be answered in order to solve those more general problems. For example, on a recent field trip in the spring the pupils in a biology class saw a number of forsythia bushes in full bloom. As it happened, all of the blossoms were on the lower branches of the bushes; none had developed on the upper branches. Several of the students noted this fact, but only two raised the question, "Why are all the blossoms on the lower branches?" These two students recognized a problem in the situation. One of the students carried his questioning still further. "Are all forsythia blossoms on lower branches? Have the lower branches been protected from the recent cold weather? Would blossoms develop on the upper branches if the plant were grown in a greenhouse?" This illustrates the ability to define some of the more specific questions which need to be answered in order to solve the more general problem.

Direct evidence of the degree of attainment of this objective

EXAMINATIONS IN MAJOR SUBJECT FIELDS

may be had by observing and recording the questions students ask when on field trips or in the laboratory. To simplify this procedure somewhat, students may be encouraged to keep notebooks in which they record problems or questions that arise in their minds as a result of their own observations and experiences at home, in the field, in the classroom, and in the laboratory.

An essay test which throws light on the ability to recognize problems is illustrated by the following:

A woman planted some flower seeds beside her house. The plants did not grow very well. The woman next door planted seeds of the same kind of flower. These plants grew very well. The first woman wondered why her flower plants did not grow as well as those of the woman next door. What information must you have before you can tell her why?

Before the situation can be explained, one must know the conditions of growth of the two groups of plants. Was there a difference in the kind of soil? Did both groups have the same intensity of light? Were both groups watered equally? Was the soil in one flower bed cultivated as well as the soil in the other? These are problems which may need answering before one can explain the difference in growth. Some further problem situations are as follows:

A small apartment house contains two upstairs apartments and two downstairs apartments. Each apartment is heated by a separate hot air furnace. To keep all apartments at the same temperature, a hotter fire must be kept in the furnaces which heat the upstairs apartments than in those which heat the downstairs apartments.

The results of wind storms in the Middle West during the spring of 1935 were different from those of previous years. Dense clouds of dust covered the area. The top soil was swept away in some places to a depth of six or eight inches. Fences

EXAMINATIONS IN THE NATURAL SCIENCES

were completely buried with the soil, and cattle could walk over them. People used dust masks made of damp cloth. People and cattle were without food.

A farmer had a flock of chickens. He noticed that some days he would get many eggs and on other days he would get very few eggs.

Students who are aware of problems will be expected to recognize and state them in situations like these. The best kind of situation is that in which the difficulty can be seen by the student but cannot be removed until further information is obtained. The student is expected to show his recognition of the problem by stating the information needed.

The evaluation of the responses in an exercise of this sort requires judgment and cannot be done as a mechanical operation. However, by directing the attention of a competent reader to the purpose of the test and suggesting a scale of values to use, some of the undesirable variations in judgments may be eliminated.

A possible test with which the teacher might experiment consists of a description of a situation and a list of questions which the students may check to indicate important problems pertaining to the situation. An example of this kind of exercise is as follows:

Directions: Below is a description of a situation. What problems or questions does this situation raise which need to be answered before you can give a reasonable explanation of it? Below the description is a list of questions which this situation might suggest. After reading the list, select and check the questions which, when answered, are likely to be most helpful in giving a reasonable explanation of the situation. Do not check more than five of the questions.

On a recent field trip in the spring, the pupils in a biology class saw a number of forsythia bushes in full bloom. As it hap-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

pened, all of the blossoms were on the lower branches of the bushes; none had developed on the upper branches.

- a. Does temperature affect the action of enzymes? (+) a.
- b. Are the blossoms too heavy for the upper branches to bear? () b.
- c. Does rainy weather after a dry spell produce blossoms on the lower branches earlier than on the upper branches? . . () c.
- d. Are the blossoms on all forsythia bushes everywhere on lower branches? (+) d.
- e. Did the buds drop off the upper branches? () e.
- f. Might the amount of rainfall affect the blossoming of forsythia bushes? () f.
- g. Does temperature affect the development of blossoms on forsythia bushes? (+) g.
- h. Have the lower branches been protected from the recent cold weather? (+) h.

If each correct answer is considered of equal value, the method of scoring discussed in previous sections may be applied here, depending upon the degree of correspondence between the results obtained by the scoring method employed and the evaluation of the direct evidence.

6. Interest in Natural Phenomena and in Solving Science Problems for Which the Student Has No Present Solution

The inquisitive behavior of pupils and the advantage they take of opportunities for solving problems and for learning about natural phenomena indicate the extent of their attainment of this objective. When a pupil asks "What happens to food when it digests?" or "How do we know that the earth revolves about the sun?" or "Why does milk sour?" it appears that he is more or less interested in natural science problems and in natural phenomena. If he asks where he can get further information about his question, he has displayed somewhat stronger evidence of his interest. If he is observed actually consulting the sources for further information, or if he

EXAMINATIONS IN THE NATURAL SCIENCES

sets up and carries through an experimental procedure to answer his question, his interest of course can be considered greater. If he behaves in this manner time and time again, and with respect to many kinds of natural science problems, the conclusions one may draw about his interest are the more reliable.

For example, a child may hear someone remark that "Thunderstorms make milk sour." Thinking that strange, he may ask where he can get further information about the souring of milk. He may forego taking part in other activities to consult these sources. He may even make an extra trip to the library. He may set up a plan by which he can find out whether thunderstorms cause the souring of milk and may carry the plan through until he has a satisfactory conclusion. The observation and recording of such incidents by the teacher are time-consuming but afford a valuable basis for judging pupil interests.

From personal interviews with children, one can get an insight into those interests which they have very little or no opportunity to express in overt behavior. Desirable rapport must be established between the child and the interviewer, so that the child will feel free to express himself verbally.

A short-cut method for getting evidence of interest in natural science problems consists in asking the student to rank a list of activities in order of interest to him. The list would include activities which show an interest in natural phenomena and also activities known to be popular with students of the age tested. The score of each activity is the rank number assigned to it. Another method consists of providing a list of activities and asking the student to rate them on a rating scale, thereby expressing the degree of interest each one has for him. The score of each activity is the value of the point on the rating

EXAMINATIONS IN MAJOR SUBJECT FIELDS

scale of interest at which it is placed by the student. Either of these methods gives a rough comparison of the student's interest in science with his interest in other activities.

7. Ability to Draw Reasonable Generalizations from Experimental Data New to the Students

Another objective of natural science teaching involves the student's ability to formulate and state — orally or in writing — reasonable generalizations which may be drawn from experimental data presented to him for the first time. This kind of behavior commonly occurs in life and is considered an important outcome of science teaching. The materials used must of course be data with which students are unfamiliar. Here again, the problem of practicability is encountered in collecting evidence desired. It is time-consuming for the students to write out their generalizations and for teachers to read and evaluate the responses. Furthermore, it usually takes three or more readers to obtain a rating which is free from individual idiosyncrasies.

A promising device which may yield results similar to the more direct evidence consists of exercises which describe the experiment, present the results obtained, and provide a list of interpretations or generalizations which students are likely to draw from the data. The list may contain four kinds of interpretations: (1) interpretations which are considered reasonable on the basis of the data obtained in the experiment, (2) interpretations which probably are true but are not entirely justified because of insufficient evidence, (3) interpretations the truth of which cannot be determined on the basis of the given data, (4) interpretations contradicted by the given data. After reading the description of the experiment and of the results obtained, the students will be expected to read each interpretation, decide which of the four kinds it is, and check it

EXAMINATIONS IN THE NATURAL SCIENCES

accordingly. An example of this type of exercise is given below.

Directions: In the following exercise, an experiment has been described. Below the description are some statements which have been suggested as interpretations of the experiment. Assume that the facts given in the description of the experiment and in the data obtained are correct; then on the basis of *these facts only* consider each statement. Mark with a figure 1 every statement which is a reasonable interpretation of the data obtained; mark with a 2 every statement which is most likely true but for which the given facts are insufficient to justify the interpretation; mark with a 3 every statement which cannot be judged as either true or false because of the insufficiency of the given facts; mark with a 4 every statement which cannot be true because it is contradicted by the data obtained in the experiment.

Nine hundred seeds of a certain plant were divided into nine groups of 100 seeds each. Each group of 100 seeds was placed in a germinator. The seeds in all the germinators were under the same conditions of air and moisture, and they were all kept in the dark. Each germinator, however, was kept at a different temperature. The various temperatures and the number of seeds which germinated within 30 days are shown in the table below.

Temperature in degrees Centigrade.....	6	8	11	13	18	25	30	35	39
Number of seeds which germinated.....	0	0	0	0	16	50	84	30	0

- a. More seeds of this variety will germinate at 28° C. than at any other temperature..... (3) a.
- b. None of the seeds germinated within 30 days at the temperatures in the experiment which were below 18° C..... (1) b.
- c. The higher the temperature, the greater the number of seeds which germinated..... (4) c.
- d. As far as the results of this experiment go, 30° C. is the optimum temperature for germinating these seeds..... (1) d.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

- e. Some of the seeds of this variety will germinate at 39° C..... (2) e.
- f. The rate at which the seeds germinated was affected by the temperature..... (2) f.
- g. Whenever seeds of this variety are germinated at 25° C., only one out of four seeds will germinate..... (4) g.
- h. A decrease in moisture content reduces the number of seeds germinating more than does a decrease in temperature..... (3) h.
- i. Beginning at 30° C., an increase of 5° C. resulted in a much greater reduction in the number of seeds germinating than did a decrease of 5° C..... (1) i.
- j. More seeds would have germinated at the lower temperatures if they had been left for a longer time in the germinators..... (2) j.

From the facts given, statements b, d, and i are reasonable interpretations. Statements e, f, and j are most likely true, although the facts are not sufficient to justify them. Insufficient facts are given to decide whether statements a and h are likely to be true or not, while statements c and g are contradicted by the data obtained in the experiment.

The kinds of statements to be marked 2 or 3 are interpretations which go beyond the data by extrapolation or interpolation. They are interpretations which include the whole population, whereas the experiment concerned only a few individuals, and interpretations which include other kinds of populations than those in the experiment.

The experimental results may be presented in various ways, such as in paragraph form, in tables, in graphs, and in pictorial descriptions.

This type of response may be varied slightly. The students may be asked to evaluate the statements in only three different ways instead of four. Reasonable interpretations which cannot be proved by the given data and interpretations which are not determinable by the facts presented may be grouped together.

EXAMINATIONS IN THE NATURAL SCIENCES

Another kind of exercise is one which describes the experiment, presents the results obtained, but provides only five interpretations. The students are expected to check with an *x* the most reasonable and most complete interpretation and to check with a *o* the interpretation contradicted by the data given. An example of this kind of exercise is given below.

Normally, frogs hibernate in the autumn by burrowing into the bottom of the pond. They emerge from hibernation in the spring. On December 10, five frogs were placed in a tank of water kept at a temperature of 55° F. All of them continued to swim in the water, and none of them burrowed into the mud on the bottom of the tank. On the same date, five other frogs were placed in a tank of water kept at 34° F. All of them went to the bottom and became inactive. On May 3, five other frogs were placed in a tank of water kept at 34° F. All of them went to the bottom of the tank and became inactive. On the same date, five other frogs were placed in a tank of water kept at 55° F. These five remained active and continued to swim around in the water.

- a. Frogs will hibernate in the autumn, but they will also hibernate in the spring. () a.
- b. The hibernation of frogs is a response to the season of the year. (o) b.
- c. The hibernation of frogs is a response to temperature.... (x) c.
- d. Frogs become inactive when they hibernate. () d.
- e. Frogs do not hibernate at 55° F. () e.

In deciding which kind of exercise to include in the test, the choice will depend upon which kind of behavior gives a good index of ability to formulate and state interpretations, upon the degree to which students are expected to discriminate between kinds of interpretations, and upon the difficulty of the exercises.

Since the objective is to teach students to interpret data which are *new* to them, unfamiliar materials must be used in the exercises. One source of these data consists of current re-

search reported in scientific journals in the natural sciences. This use of new research data precludes the possibility of parrot-like repetition of memorized interpretations and forces the students to formulate their own. Some students can repeat from memory interpretations which were made for them but cannot interpret new data for themselves. With other students, the reverse is true.

In tests in which the students state, orally or in writing, their own interpretations of new data, the responses may be evaluated on a scale of quality ranging from very poor (numerical value zero) to very good (numerical value 10).

A difficulty encountered in evaluating essay responses or observed behavior is that inadvertently, in judging the responses, the grader will consider other objectives rather than those which represent the purposes of the evaluation. For example, a student who writes long answers, or whose handwriting is good, or who uses the English language fluently may receive a high grade on a very poor interpretation. A student who presents a large amount of information but offers poor interpretations is sometimes given high ratings. Such students as these camouflage the issue, and some graders do not penetrate the "smoke screen." A valuable procedure for overcoming this difficulty is to provide a set of specifications for the graders to study and use in evaluating the responses. A specimen set is presented below.

Specifications for Grading

1. Grade the first question on *all* the papers first, then the second question on all the papers, and so on.
2. The best answer is one which gives the most reasonable interpretation, i.e., the interpretation which is as complete and as broad as can be made without going beyond the facts given in the description of the experiment and in the results obtained.

3. After reading each answer, judge how well (to what degree) it approaches the best answer, and in accordance with this judgment assign the answer a number on the scale of quality.
4. Note that the points on the scale of quality are equally distant from each other. The distance between an answer having a value of 6 and one having a value of 7 is the same as the distance between an answer having a value of 3 and one having a value of 4, and so on. Try to place the students' answers on this scale of equal intervals.
5. Do not judge the answers from the standpoint of spelling or English usage. Judge the answers only on the degree to which they are as complete and as broad interpretations as can be made without going beyond the facts given in the description of the experiment and in the results obtained. The purpose of the grading is to evaluate the students' ability to make reasonable interpretations and not to evaluate their use of English.

8. Ability to Plan Experiments and to Test Hypotheses

What kind of evidence shows that students can plan experiments for testing promising hypotheses? Observation of the behavior of students in the laboratory while they are formulating plans for an experiment and demonstrating how a given hypothesis may be tested will afford some evidence. Oral interviews and written essay tests consisting of such questions as, "How can one find out whether a certain muscle in an animal is an extensor and not a flexor? — How can one find out whether the diffusion of oxygen gas through a membrane increases or decreases with increases of pressure? — How can one tell whether light affects the germination of corn seeds?" are methods by which direct evidence can be collected. Students are expected to state the plans for the experiments to test these hypotheses. The hypotheses used should not have been previously demonstrated or discussed in class or in the textbook. They must be new to the students, since the exam-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

ination should not be a test of the student's memory of a specific experiment proposed by some one else.

A possible test consists of exercises in which several methods of testing the hypothesis are provided. The students may decide which is the best method and check it.

How can one find out that a certain muscle in a frog's hind leg is an extensor and not a flexor?

- a. Stimulate the muscle with an electric current and watch the movements of the muscle as it contracts and relaxes..... () a.
- b. Remove the skin from the hind leg of a freshly killed frog. Pass an electric current through the muscle. If the leg kicks out, the muscle is an extensor muscle. () b.
- c. Suspend the hind leg of a freshly killed frog so that the leg is free to move. Stimulate the muscle with an electric current. If the leg extends and does not bend when the muscle contracts, the muscle is an extensor and not a flexor..... (+) c.
- d. Suspend the hind leg of a freshly killed frog and remove the skin. Stimulate the muscle, taking care not to stimulate other muscles. If the leg moves, the muscle is an extensor..... () d.
- e. Remove the skin from the hind leg of a freshly killed frog. Clamp the leg in position so that it is free to move. Stimulate the muscle with an electric current from a small battery. If the muscle contracts, it is an extensor muscle..... () e.

A second possible exercise consists of a hypothesis and two lists of statements. The first list contains statements of kinds of evidence, and the student checks those which he believes must be shown in order to test the hypothesis. The second list contains statements of procedures by which the evidence may be collected. An illustration of this kind of exercise follows:

EXAMINATIONS IN THE NATURAL SCIENCES

How can one find out that a certain muscle in a frog's hind leg is an extensor and not a flexor?

It would need to be shown that:

- a. The muscle relaxed..... () a.
- b. The leg did not bend when the muscle contracted..... (+) b.
- c. The leg moved when the muscle contracted..... () c.
- d. Other muscles which were not stimulated did not extend the leg..... () d.
- e. The leg extended when the muscle contracted..... (+) e.
- f. The muscle is a striated muscle..... () f.

Procedures which would need to be used:

- g. Tie the ends of a muscle dissected from the hind leg of a freshly killed frog to the ends of a hinge..... () g.
- h. Suspend the hind leg of a freshly killed frog so that the leg is free to move in both directions..... (+) h.
- i. Stimulate the muscle with an electric current..... (+) i.
- j. Examine the dissected muscle under a microscope..... () j.

9. Ability to Apply Scientific Generalizations to New Situations

Still another objective involves the application of past experience in solving new problems. Experiments have been performed again and again resulting in the same generalizations. For example, it has been demonstrated with many gases that the volume increases as the temperature increases if the pressure is kept constant. By applying this principle, accurate predictions can be made of what will happen when the temperature of a particular gas is raised and the pressure kept constant. There are many principles and facts in the natural sciences which can be used thus in predicting the outcome of a new problem situation. Evidence of the ability to apply scientific generalizations may be obtained by presenting the student with an unfamiliar problem situation and asking him to state what will happen and to explain his prediction by citing the principles and facts which he applied to the case. The following are examples of this kind of question.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

A full-grown normal rabbit which has been fed on carrots and water for a week is changed to a diet of beans, clover, and water. What change will occur in its kidney excretions in eight hours, and why?

Two identical cubes of wood are sterilized and placed in separate moist chambers. Block A is kept sterile. Block B is inoculated with a fungus, which grows and thrives in the wood. How will the dry weight of these blocks probably compare at the end of three months? Why?

A promising type of exercise which may give a good index of the desired behavior consists of a statement of the problem, a list of predictions which students are likely to make, and a list of facts and principles by which each student may indicate the line of reasoning followed in determining his prediction. The latter list will include both true and false statements. After the student has checked the outcome or outcomes which he thinks will probably occur, he then selects and checks in the second list the facts and principles which he applied in his prediction. An exercise of this kind is as follows:

Directions: In the exercise below, a problem is presented. Below the problem are two lists of statements. The first list contains statements which can be used to answer the problem. Place a plus sign (+) in the parentheses after the statement or statements which tell *what will probably happen*. The second list contains statements which can be used to explain the right answers. Place a plus sign (+) in the parentheses after the statement or statements which *give the reasons for the right answers*. Some of the other statements are true but do not explain the right answers; *do not check these*. In doing these exercises then, you are to place a plus sign (+) in the parentheses after the statements which *answer the problem* and which *give the reasons for the right answers*.

Problem

Coal gas which has not been mixed with air previous to burning is burned at a gas jet. At another similar jet the coal gas is

EXAMINATIONS IN THE NATURAL SCIENCES

mixed with air before it is burned. Will there be any difference in the amount of light given off by the flames of the two gas jets? Why? If a cool aluminum pan is placed over each flame will there be any difference in the amount of soot deposited on each pan? Why?

Predictions

The flame at the first gas jet will give off:

- a. More light than the flame at the second gas jet. (+) a.
- b. The same amount of light as the flame at the second gas jet. () b.
- c. Less light than the flame at the second gas jet. () c.

The soot deposited by the first gas jet will be:

- d. More than that deposited by the second gas jet. (+) d.
- e. The same in amount as that deposited by the second gas jet. () e.
- f. Less than that deposited by the second gas jet. () f.

Reasoning

From the following statements select and check the ones which indicate the line of reasoning you followed in making your predictions above.

- g. Incomplete combustion leaves some uncombined carbon in the flame. (+) g.
- h. The presence of nitrogen retards combustion. () h.
- i. Particles of uncombined carbon glow when heated. (+) i.
- j. Combustion is more complete in the first flame. () j.
- k. The first flame contains very little if any uncombined carbon. () k.
- l. The amount of air mixed with the gas does not affect the amount of light produced by the burning gas. () l.
- m. Uncombined carbon in a flame is deposited on a cool surface placed in the flame. (+) m.

In the above exercise, statements which answer the problem are a and d. Statements g, i, and m are reasons which help

EXAMINATIONS IN MAJOR SUBJECT FIELDS

to explain why a and d will probably happen. It will be noticed that statement h is a true statement but does not give a reason for any of the right answers; the other statements are not true and do not explain the right answers.

Another type of exercise, requiring only the application of facts and principles known to be true, may consist of several problems, a list of predictions, and a list of true statements of facts and principles. Students may indicate the predictions for each problem and the facts and principles which they think apply. The students know that all the statements of facts and principles are true and that they are to select only those which help to explain their prediction or predictions for each problem. An exercise of this kind is given below.

Directions: Below are two problems. After the problems are two lists of statements. The first list contains statements which can be used to answer each problem. Place the problem's number after the statement or statements which *answer that problem*. The second list contains statements which can be used to explain the right answers. Place the problem's number after the statement or statements which *give the reasons for the right answers*. Some of the other statements are true, but do not explain the right answers: *do not use these*.

1. Some hydrogen generated from a water solution of sulfuric acid was passed through a coil surrounded by liquid air. What will probably be the effect upon the gas and the inside of the coil? Why?
 2. Some solid platinous oxide, PtO , is heated to a temperature of 800°C . in a closed tube. What will probably be found in the tube if it is examined after the heating? Why?
-
- a. Nothing will happen. — — a.
 - b. The gas in the tube will ignite a glowing splint. — ² b.
 - c. The hydrogen will be dried. ¹ — c.
 - d. A liquid will be found in the tube. — — d.

EXAMINATIONS IN THE NATURAL SCIENCES

- e. Ice will be found upon the inside of the tube..... 1 — e.
 f. The residue will appear different from the original
 solid..... — 2 f.
 g. Steam will be found in the tube..... — — g.

From the following statements select and number the ones
 which indicate the line of reasoning you followed in making
 your predictions for each problem.

- h. The activity series for metals arranged from the
 most active to the least active is: calcium, sodium,
 zinc, lead, hydrogen, mercury, silver, platinum, and
 gold..... — 2 h.
 i. Gases generated in water solutions carry water
 vapor..... 1 — i.
 j. Molecules of gases above their boiling points have
 such velocities that their attractions for each other
 are practically negligible..... 1 — j.
 k. Oxides of metals below mercury in the activity
 series are decomposable at temperatures above
 600° C..... — 2 k.
 l. Molecules of a vapor exercise cohesive attraction
 for each other at temperatures below their boiling
 points..... 1 — l.
 m. Ordinarily an oxide is different in appearance from
 its constituent metal..... — 2 m.
 n. The molecular theory assumes that when a pure
 substance is cooled below its freezing point cohesion
 between its molecules forms crystals..... 1 — n.
 o. Oxygen gas is an active supporter of combustion... — 2 o.
 p. Metals below hydrogen in the activity series do not
 react with acids..... — — p.
 q. At higher temperatures the velocities of molecules
 may become so great as to knock the constituent
 atoms free from each other..... — 2 q.
 r. Atoms of some elements group into pairs and form
 molecules of that element..... — 2 r.
 s. Hydrochloric and acetic acids are non-oxidizing
 acids..... — — s.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

Direct evidence, obtained from essay tests in which students formulate and state the predicted outcomes of problem situations and the chain of reasoning they used in making the predictions, may be evaluated by using a scale of quality similar to that described in previous sections. Another method of scoring such tests is to indicate in a set of specifications the predictions and the facts and principles for which credit may be given, together with the numerical amount of credit to be given to each. A score chart may be arranged on which the graders can record their judgments. The student's score on the problem may be considered as the sum of these values. Other possible methods of scoring are similar to those described previously. Still another method is to select the minimum combination of statements which include an important prediction and a logical chain of reasoning. A certain number of points credit may be given only if the student has checked or presented this minimum pattern of response. Further credit may be given for other contributing ideas checked or presented in addition to this minimum pattern.

10. Skill in Laboratory Techniques

The various skills involved in the manipulation of laboratory apparatus and in the performance of laboratory experiments — for example, skill in using the microscope, in making dissections, in making mounts, in setting up chemical or physical apparatus, and in making analyses of unknown chemicals — constitute another objective. Direct evidence of achievement in these skills may be obtained by observing the student at work and checking and evaluating each step of the work as he proceeds, or by evaluating the product of his work.

A check list is helpful in evaluating the specific activities which the students perform while they are working. Skillful workers may be observed and each important activity recorded

EXAMINATIONS IN THE NATURAL SCIENCES

in developing a key list of activities involved in the manipulation. These activities may be arranged in the order in which skillful workers perform them. A student's skill in dissecting the muscles of a frog's hind leg, for example, may be evaluated by checking against this list the activities which he performs satisfactorily. This is an individual procedure and involves the observation of one student at a time.

A second procedure is to have all of the students make the dissection and to evaluate the product of each dissection immediately after the students have finished by checking against a list of characteristics which the dissection should show if properly executed. Those students whose dissections are very poor can then be observed individually as they repeat the work and their errors in procedure noted. Each student may be given a record of those activities which he has performed satisfactorily and those which he has not. From this he learns in what activities he needs to develop more skill.

A set of directions to the student and a check list for the instructor which may be used in evaluating skill in dissecting muscles of the frog's hind leg are given below. The numbers in the parentheses indicate the amount of credit which may be given if the dissection has the desired characteristic or if the activity was performed satisfactorily. A score on the product and a score on the procedure may be obtained as indicated at the bottom of the respective columns.

Directions to student: On the table you will find a supply of frogs and the materials needed to make a dissection.

- (1) Remove the skin from the frog's hind leg.
- (2) Dissect the gastrocnemius and the tibialis anticus longus muscles of the lower leg free from other muscles but left attached to the bones, showing their origin and insertion.
- (3) As soon as you have finished, notify the laboratory instructor so

EXAMINATIONS IN MAJOR SUBJECT FIELDS

that the condition of the dissection can be scored before it has deteriorated.

You will be allowed *exactly* twenty minutes to make this dissection.

Check List for Instructor

Scoring of Dissection

- a. Is the skin completely removed from the leg and foot? (1) _____ a.
- b. Are the muscles, tendons and joints uninjured and intact? (1) _____ b.

Score the remaining items for the gastrocnemius muscle and the tibialis anticus muscle separately. Allow the specified number of points credit for *each* muscle.

Gast. Tib.
Ant.

- c. Is the muscle completely separated from adjacent muscles? (2) _____ c.
- d. Is the muscle attached at the origin? (1) _____ d.
- e. Is the muscle fully dissected at the origin, its attachment distinct? (1) _____ e.
- f. Is the muscle attached at the insertion? (1) _____ f.
- g. Is the muscle fully dissected at the insertion, its attachment distinct? (1) _____ g.
- h. Is fascia of muscle smooth, not torn? (2) _____ h.
- i. Are the fiber bundles entire, not frayed out? (1) _____ i.

Score on Dissection

Sum of items a to i inclusive _____

Scoring the Procedure

- j. Does the student hold or place the frog so that its head is away from the student? (1) _____ j.
- k. Does he handle the frog with his hands when it expedites matters to do so? (1) _____ k.
- l. Does he cut through skin around the thigh near the body? (1) _____ l.

EXAMINATIONS IN THE NATURAL SCIENCES

- | | |
|--|------------|
| m. Does he remove all the skin from the leg and foot? | (2) ——— m. |
| n. Does he remove all skin in less than 5 minutes? | (1) ——— n. |
| o. Does he separate gastrocnemius muscle from adjacent muscles? | (2) ——— o. |
| p. Does he separate tibialis anticus longus muscle from adjacent muscles? | (2) ——— p. |
| q. Does he use blunt end of scalpel, or probe, or pencil point or fingers to separate the muscles, and not the sharp point of an instrument? | (2) ——— q. |
| r. Does he avoid cutting or tearing fascia of muscle? | (1) ——— r. |
| s. Does he avoid cutting muscles free from their points of origin? | (1) ——— s. |
| t. Does he avoid cutting muscles free from their insertion? | (1) ——— t. |
| u. Does he avoid removing muscles from the leg? | (1) ——— u. |
| v. Does he avoid shredding or tearing muscle bundles? | (1) ——— v. |
| w. Does he finish entire dissection in 15 minutes or less? | (1) ——— w. |

Score on Procedure

Sum of items j to w inclusive

—————

It is also possible to use a scale in evaluating each characteristic of the product or each activity in the procedure. When the amount of credit which may be given is one point, the scale is from 0 to 1, in which 0 represents an unsatisfactory performance while 1 represents satisfactory performance. When the amount of credit which may be given is two points, the scale is from 0 to 2. The latter scale permits an intermediate value.

Sources of Materials for Examinations in the Natural Sciences

In the preparation of any adequate examinations designed to obtain evidence of the degree to which the objectives of science

courses are being attained a reservoir of basic materials is needed. This reservoir should include a list of important facts, technical terms, and principles which students might be expected to remember; a list of common misconceptions which the course might help to eliminate; a list of sources, both reliable and unreliable, of scientific information; collections of problems, problem provoking situations, and problems to be analyzed; experimental data or facts encountered in every day life from which students should be able to draw generalizations; a list of hypotheses which could be tested by the students; a list of scientific principles which the students should be able to apply to new situations; and a collection of laboratory techniques which the students are expected to master. The actual content of the examinations for each of the objectives may then be obtained by selecting appropriate samples from these lists.

For convenience and flexibility, each fact, principle, problem, or misconception may be written on a small card. A hypothesis which is to be tested might be written at the top of the card and below it the questions which need to be answered to check the hypothesis. The cards on which scientific generalizations or principles are listed should also include problem situations new to the students which provide opportunity for the application of these principles. Each laboratory technique included in the examination materials will include detailed analysis of the steps involved and of the product resulting. If the cards are filed under the topics of the course, it will permit easy revision, addition, or subtraction, to correspond with changes in the course from year to year.

In preparing these lists, the teacher will usually turn first to his syllabus or course of study. In some cases, curriculum studies have been made which give evidence of the frequency with which some of the facts, principles, or problems are en-

countered in various activities of life, or of other factors which help to indicate their probable significance. Such studies are helpful in selecting items for the list of important information. In building some of the Cooperative science tests, lists of generalizations have been submitted to a large number of science teachers for their judgments of the significance of each generalization to the student. These lists will frequently prove helpful to the teacher in compiling his own. An analysis of appropriate textbooks and reference books will reveal additional items which, on careful consideration may prove to be important. Reports of current research, including the popular summaries prepared by Science Service, also frequently present interesting and significant problems for further investigation.

Perhaps one of the most useful sources for examination materials is the science teacher's notebook. In this the teacher will record his observations of the local environment, and of interesting data which the students meet in everyday life. He may record here problems or misconceptions which are brought out in class discussions, interviews with students, written papers, and previous examinations. Newspapers, magazines, popularized books on science, and adult discussions in the community frequently reveal misconceptions, problem situations, or untested hypotheses. The habit of recording such instances is helpful in building the basic lists of examination materials.

Often it is helpful to have several other competent teachers go through the tentative lists obtained by these methods and indicate their judgments of the value of the items to the students. In actually building the examinations for the different objectives, the lists should be adequately sampled. The problem of getting a satisfactory sample is treated in the chapter on "Construction of Tests."

EXAMINATIONS IN MAJOR SUBJECT FIELDS

PROCEDURES FOR DEVELOPING A COMPREHENSIVE PLAN OF MEASUREMENT TO COVER THE ENTIRE FIELD

Planning of Measurement Devices to Cover All Major Objectives

A comprehensive program of measurement includes various devices designed to collect evidence of achievement in all the important objectives of natural science teaching. High achievement in one objective does not necessarily mean high achievement in another; nor does low achievement in one necessarily mean low achievement in another. Students who have acquired a large fund of information, for example, frequently do not have the same degree of success in interpreting new experimental data.

The first problem in developing a comprehensive program of measurement is to formulate the objectives or expected outcomes of the course. The second problem is to clarify the objectives by defining them in terms of changes expected in the behavior of students. These two steps are described in some detail in an earlier chapter. The third problem is to make a collection of situations in which the behavior may be expressed.

A number of questions must be answered in considering methods for collecting evidence of achievement in each of the important objectives of teaching. What are the conditions under which the behavior is expected to take place? Will the examination be "sprung" or will the students be notified ahead of time? Will they be directed to guess or will they be directed to avoid guessing and informed why? Will the problem situations be new to the students or will they be familiar ones? Will they be "paper and pencil" situations or will they include laboratory, library, classroom, and free-time situations? Will the students have time to finish the examination or will there be a time limit? Will they be permitted to consult sources of information during the examination? A guiding principle to

follow is to consider the purpose of the method and the factors to be controlled in the situation. For example, if one wishes to know what information students can recall at a certain time without reviewing the material immediately before the evidence is collected, the students will not be notified ahead of time and the examination will be "sprung."

Distribution of These Different Measurements During the Instructional Period

If all evidence of achievement in a comprehensive program of measurement were collected at one time, a very long time would be required to obtain sufficient evidence on which to base reliable judgments. There may not be sufficient time to devote to many examinations at the end of each month, six weeks, quarter, semester, or year. At present not many of the needed practicable examinations which give a good index of the more direct behavior expressing each objective are available. However, the collection of evidence may be distributed over the instructional period. Paper-and-pencil examinations of information acquired, of the understanding of terminology, of the ability to interpret new data and to apply principles may be given periodically during the course. One or more laboratory periods may be devoted to measuring laboratory skills.

Evidence of interest in solving problems in natural science and interest in natural phenomena may be observed in the laboratory, classroom, interviews, and other situations from time to time. It is important that anecdotal records of the students' behavior be made, i.e., notes of significant pupil reactions, for evaluation at a later time by the teacher and by others. These records constitute evidence of achievement in the objective. It is well to keep the record of behavior free from interpretation of the behavior. The behavior and the conditions under which the reaction took place should be de-

scribed objectively, so that others who read the record will get an accurate picture of the conditions and behavior of the student. The record should be a substitute for actual observation to persons who were not present when the behavior took place and who did not know the events leading up to it. The interpretation of this behavior record may be made on the reverse side of the card. Thus it is possible to obtain from each individual an independent interpretation of the behavior. Written papers on topics giving references to reliable sources of information may be collected from time to time. The student's ability to identify structures of plants and animals and their functions may be noted in the laboratory during class periods and on written examinations.

As practicable examinations are developed it will be possible to collect more economically much of the evidence about the various kinds of changes in students which are important outcomes of the course.

Cooperative Examination Building

The development of a comprehensive program of examinations is a large task for one teacher to assume and to accomplish in a short time. Cooperative attack on the problem by teachers of the same school system or of different school systems is a more promising method of developing examinations. Periodic meetings of natural science teachers may be held for this purpose. Each teacher may formulate and pool his objectives with those of the other teachers. There may be differences of opinion about objectives, but they usually are only disagreements on degrees of importance. Separate examinations for the various objectives will enable each teacher to test for those outcomes which he considers most important and to weight each objective in accordance with his judgment of its value.

After the objectives have been formulated and clarified by indicating the kind of student behavior involved, each teacher may choose one or more for which he will be responsible in devising methods of measurement. His plans may be presented to the group for criticism and suggestion, the other teachers cooperating wherever he needs their assistance. By this plan, all teachers participate in developing test materials for all the important objectives, but each teacher assumes the responsibility for only one or a few of those objectives.

When it is inconvenient for teachers to meet and work together, one teacher working alone may choose one objective for which he will develop examinations during the year. The following year another objective may be chosen, and so on.

The Cooperative Test Service, which is a non-profit agency organized under the auspices of the American Council on Education to produce and distribute tests, is following such a cooperative procedure in developing science tests. Teachers in various schools and colleges are working upon tests for each of the science objectives. With the assistance of test technicians they are experimenting to discover practicable methods of testing certain objectives which are difficult to test directly. Through such cooperative efforts a comprehensive program of examinations becomes possible. It is essential that our examining procedures become more comprehensive. Some of the most significant objectives of science teaching are not commonly tested. If science examinations are to have their greatest value to student and teacher alike, they must give evidence with reference to all of the important outcomes of courses of science.

QUESTIONS FOR DISCUSSION

1. In talking with his general science teacher, a boy said, "This morning you were telling us about being sensitive to general science problems in our environment. What does that mean?"

EXAMINATIONS IN MAJOR SUBJECT FIELDS

Maybe I would like to be like that." What can the teacher tell him? Would the explanation be useful in constructing a test? Would it be a paper and pencil test?

2. At the science section meeting of the State Education Association, a teacher heard much talk about the appreciation of science as an objective of science teaching. Being interested in the development of the boys and girls in his classes he wanted to find out whether they are learning to appreciate science. How could he do this? Comment briefly on the steps he could follow.
3. A teacher ordered a sample copy of a biology test. When it arrived, he showed it to one of his colleagues and said, "I now have the test, but how can I tell if it is any good? On what basis should I judge this test?"
4. A general science teacher thinks that one of the things his courses should do for the children is to help them to develop the habit of asking for evidence to support important statements presented to them. Can he measure this objective? If so, how?
5. A boy's library card for the month of April shows that he has withdrawn 32 books. On closer examination of the titles, it is found that all of them are about chemistry, physics, and biology. Does this mean anything? Is it a test?
6. One of the objectives of the general science class is to understand one's environment. The teacher wanted to find out whether his children were learning to understand their environment. What evidence could he collect which would show this? How could he collect the evidence?
7. A teacher worked diligently on an examination for evaluating his students' ability to test promising hypotheses in biology. After he had prepared the examination he showed it to a fellow-teacher who remarked that it was only another objective test. He added, "The way to test this outcome of teaching is an essay test. You can't use an objective test." What do you think of this criticism? What evidence would you need before you would accept the objective test for use in your classes?
8. Before bringing to the attention of his biology class the prob-

EXAMINATIONS IN THE NATURAL SCIENCES

lem of eugenics, the teacher wanted to find out what attitudes the class had toward the problem. He prepared a list of statements representing various degrees of attitudes. He asked each student to check each statement showing that he agreed, disagreed or was uncertain about the statement. He wanted the attitude of the class toward this problem to be expressed numerically on a continuum. How might he go about doing it?

9. At a meeting of Eastern High School faculty, several of the teachers expressed an interest in improving their examinations. They discussed the possibilities and finally felt that the job was too big. One man characterized the feeling of the group when he said, "I am sure I won't find time to work on examinations for all of the important objectives in my science classes. The job is overwhelming." What practicable plans can you suggest to these teachers?
10. A girl made the highest score in the typical true-false test in physics given at the end of the semester. In fact she usually made high scores on these tests during the semester. Her teacher was perplexed because so often this girl could not explain the simplest phenomena in her environment which were not discussed in class or in the textbook. What problems in evaluation does this situation raise?
11. A boy came home from school the last day of school and at dinner that evening broke the sad news to the family that he failed in chemistry. His father asked, "What does that mean?" The boy replied that he did not pass. He didn't get a high enough grade. He would have to take the course again next year. "Yes, but how can they tell when you fail or not?" asked his father. "By my marks, I guess," said the boy. Further discussion brought out nothing more. What important question or questions about examinations does this situation raise?

CHAPTER VI

EXAMINATIONS IN THE FOREIGN LANGUAGES¹

I. LATIN

OBJECTIVES

THE *Report of the Classical Investigation*² contains the most extensive discussion of the aims of teaching Latin — their validity, the course content and the procedure by which they may be attained — that has appeared since the methods of investigation employed in the natural sciences have been applied to educational problems. As a result of the inquiry, the committee responsible for the report formulated conclusions in respect to objectives substantially as follows:³

¹ The committee responsible for the preparation of the chapter dealing with examinations in foreign languages consists of:

Algernon Coleman, University of Chicago, *Chairman*
Harry E. Ford, University of Toronto
V. A. C. Henmon, University of Wisconsin
James B. Tharp, Ohio State University
B. L. Ullman, University of Chicago

The Latin section of this chapter was read in manuscript by W. L. Carr of Teachers College, Columbia University, John Seddon Fleet of Culver Military Academy, and Laura B. Woodruff of the Oak Park (Illinois) High School; the section on modern languages by Edna C. Dunlap and Millicent Stebbins of the Parker High School, Chicago, Helen M. Eddy of the State University of Iowa, and Peter Hagboldt of the University of Chicago. J. M. Stalnaker of the Board of Examinations of the University of Chicago read both sections in manuscript, as did the chairmen of the various committees responsible for this volume. The Committee is grateful for much valuable advice and criticism from these sources, but assumes entire responsibility for all defects in the form and in the content of the chapter as it stands.

² Part I. Princeton University Press, 1924.

³ *Ibid.*, p. 78.

A. Primary

The indispensable primary immediate objective is progressive development of ability to read and understand Latin.

B. Ultimate

1. Increased understanding of the elements in English related to Latin; increased ability to read, speak, and write English; development of literary appreciation.

2. Increased ability to learn other foreign languages.

3. Development of a historical and cultural background.

4. Elementary knowledge of the simpler principles of language structure.

5. Development of correct mental habits and of right attitudes toward social situations.

Later in the report,⁴ we note the addition of the words "as Latin" to the formulation of the immediate or primary aim. As early as 1899 the Committee of Twelve of the American Philological Association had declared, "The student should be carefully trained to take in the meaning of the sentence in the order in which it stands, before translating it."⁵ Reaffirmed, as we have seen, in the Report of the Classical Investigation, the principle involved is accepted in a subsequent report of a committee on Latin of the North Central Association of Colleges and Secondary Schools.⁶ In another similar document of great importance,⁷ however, this view of how Latin should be "read" is ignored. Indeed, in this report

⁴ *Ibid.*, p. 188.

⁵ *Proceedings*, vol. xxx, 1899. See also the *Report of the Committee of Ten of the National Education Association*, 1894, pp. 70-71.

⁶ *High School Curriculum Reorganization*. Ann Arbor, Michigan: North Central Association, 1933, pp. 148-174.

⁷ *Report of a Study of the Secondary Curriculum*. Milton, Massachusetts: The Secondary Education Board, 1932, pp. 84-99. See especially pp. 84-85.

the implications are very clear that the normal and desirable procedure by which an American pupil understands a Latin sentence and demonstrates his understanding is to translate it. Whatever may be the merits of the two doctrines in respect to the way pupils should learn to read Latin, it is highly improbable that the doctrine enunciated by the Committee of Twelve of the American Philological Association in 1899, and reiterated vigorously in the Report of the Classical Investigation in 1924, has been widely accepted in the classroom, for there are only scanty traces of its effect on examinations in Latin.

The College Entrance Examination Board took cognizance fairly promptly of the recommendations of the Report of the Classical Investigation and made certain changes, both in its statement of requirements and in the examinations that have been set, beginning with June, 1929. The change affected especially the examination for the third and fourth years, and consisted chiefly in inserting a Latin passage followed by content questions to be answered in English, and in introducing a few questions of a "cultural" nature. Specimens of the latter may be taken from the June, 1933, examination, based on a passage from Cicero's *Philippics* and a passage from Ovid's *Metamorphoses*, both of which were set for translation. The questions are:

1. Explain the derivation, stating the meaning of all the Latin elements of which each word is composed: temporal, averse, invincible.
2. Against whom were Cicero's *Philippics* delivered? At what time in Cicero's life? With what consequences to Cicero?
3. What were the characteristic functions of Apollo?
4. Describe some other significant apparition of a god in the Latin poetry you have read.

The two-year examination, however, consists of: a passage to be translated into English; seven grammatical questions containing 31 items, which involve declension of nouns and pronouns, conjugation of verbs, comparison of adjectives and adverbs, an explanation of uses of cases, tenses, and moods; three English sentences to be translated into Latin. The examination for 1934 follows the same pattern. One "three-year" combination (1933) calls for: a translation into English of a passage from Cicero; the "explanation" of four grammatical phenomena; the first and second "cultural" questions given above; content questions on a passage from Cicero, *In Verrem*, the replies to be written in English; and an English passage of about one hundred words to be translated into Latin. There are other combinations for "three-year" and several of the same sort for "four-year" Latin, but all these vary in the source of the reading passages rather than in the type of examination. The 1934 examinations conform to the same pattern. Evidently, then, College Board examiners expect two-year Latin students to know the forms of the various parts of speech, to "explain" their uses, to translate into English passages from such writings as *De Bello Gallico*, and to turn complex English sentences into Latin. Since only about 17 per cent of those who begin Latin in public secondary schools continue the subject for more than two years in school, and since only a portion of those who leave school after two years of Latin continue the subject in college,⁸ the majority of those who begin Latin must derive in two years whatever educational benefits their classroom study of the subject may be expected to produce. Consequently,

⁸ Algernon Coleman, *Teaching of Modern Foreign Languages in the United States*, Publications of the American and Canadian Committees on Modern Languages. New York: The Macmillan Co., 1929, pp. 21 (Table I, Latin), 26. *Report of the Classical Investigation*, p. 31.

although the pupils who do well on such examinations as those set by the College Board may have derived these benefits, the answers of the two-year group can prove at best only that they are capable of translating a given Latin passage into English, of giving certain forms, of recognizing others and grouping them under the proper categories, and of making a Latin version of some English sentences.⁹

In view of the insistence, throughout this volume, on the unescapable relationship between the aims of a course of study and the examinations set to measure progress toward those aims, it is appropriate at this point to consider which of the immediate and ultimate aims listed in the *Report of the Classical Investigation* and incorporated, as we have seen, in other important statements, we may reasonably expect to be reflected specifically in the examinations set by teachers of Latin.

⁹ The January, 1934, "New-Type" examination of the New York Regents for two years of Latin consists of two passages to be translated into English (40 points), seven questions on syntax (10 points), a passage of 73 running words on which seven questions are based to be answered in English (10 points), four sentences to be translated into Latin (16 points), eight items to be conjugated, declined, etc. (10 points) six English cognates to be "defined" and associated with their Latin etymons (6 points), four questions of a historical nature bearing on Caesar's life and campaigns (8 points). The corresponding "Old-Type" examination seems to differ from the above in procedure only in presenting three Latin passages for translation (48 points). The "New-Type" examination for three years follows the same procedure and contains similar items; the fourth-year "New-Type" consists of two Latin passages for translation (70 points), four questions of syntax and on versification (3 points), four questions on the content of the *Aeneid* (8 points), five quotations to be translated into English (5 points), and a quotation to be given from memory (4 points).

The study by Crawford and Burnham (*School and Society*, vol. xxxvi, 1932, pp. 344-352, 378-384) reveals a very low correlation between scores on the C.E.E.B. examinations in Latin and first-semester grades at Yale. The correlations based on the scores of 286 students are .22 for the class of 1933 and .34 for the class of 1934. As the reliability coefficient of the Yale examination was .74 these entrance tests in Latin had little predictive value. Such a piece of evidence constrains us to reflect seriously on the content and the form of examinations which affect so notably teaching and examining in Latin the country over.

Properly constructed examinations should provide an adequate measure of the progress made by pupils in ability to read Latin; knowledge of English derivatives from Latin and of their meanings; and growth in historical-cultural background as a result of the study of Latin. It should even be possible for the teacher to determine, in the case of each individual pupil, whether he reads Latin in the order in which it is written, although a written examination can throw little light on this point.

In respect to the other objectives entered on the list given above, we already have some evidence of the validity of the claim that students who succeed in Latin make better progress in a second foreign language than equally able students who lack this experience.¹⁰ An objective of this type, however, can have no specific effect on the construction of examinations in Latin. It is also unlikely that the Latin examination can, at the secondary school stage, provide the opportunity to test improvement specifically in reading, speaking, and writing English, and in literary appreciation. Furthermore, we can hardly venture to believe that examinations in Latin will throw much light on the pupils' growth in "mental discipline" and in right social attitudes. But in addition to testing reading ability, with its concomitant knowledge of vocabulary, forms, and syntax, the Latin examination, as has been asserted, can measure progress toward the other aims, as they have been enumerated. Whether a list of objectives should include types of attainment that do not lend themselves to any direct evaluation in degree, is a question that lies beyond the limits of the present discussion.

The College Board examination under consideration casts

¹⁰ T. J. Kirby, "Latin as a Preparation for French," *School and Society*, xviii (November 10, 1923), pp. 563-569; L. E. Cole, "Latin as a Preparation for French and for Spanish," *ibid.*, xix (May 24, 1924), pp. 618-622.

no light on the pupils' progress toward the goal of reading Latin "as Latin." Nor does it, at the two-year level, test progress in knowledge of the Latin elements in English or reveal any development of a general historical-cultural background, modest as progress in these directions must inevitably be at this stage. Only the examinations at the more advanced stages take these aims into account. The examinations of the Secondary Education Board follow the general pattern of the two-year examination of the College Board, which, in all likelihood, is representative of Latin examinations in most secondary schools. If, therefore, examinations and aims are, in actuality, closely allied, we may conclude that in the minds of most teachers the immediate aims of the two-year course, as reflected in current examinations, are: the possession of a very limited English-Latin vocabulary and of a somewhat larger Latin-English vocabulary; a knowledge of the commonest forms of the various parts of speech and of their uses; the ability to use this vocabulary and knowledge in translating sentences and passages into and from Latin.

The following paragraphs provide a discussion of ways of improving and of enlarging the scope of Latin examinations in general.

EXAMINATIONS

1. Vocabulary

As a test of progress in the powers enumerated in the preceding paragraphs, the second-year examination referred to would be much more effective if it contained specifically an objective test of growth in vocabulary. Investigators in the field of modern language teaching have noted that scores in vocabulary correlate better with reading scores than do scores

in grammar. The study by Haage¹¹ shows a definitely higher correlation for all four years between vocabulary and comprehension than between forms and comprehension. Thus the scores on vocabulary provide indirectly a useful index of progress in reading ability. Such a test has all the more validity in Latin, as compared with modern languages, since the College Board word list has, for nearly a decade, provided a definite and widely accepted basis for vocabulary study. In modern languages, similar uniformity is being approached only by degrees and with some opposition on the part of a good many members of the profession.

A useful technique in constructing a vocabulary test is the multiple-choice technique. A Latin word is accompanied by, for instance, four or five English words, one of which is an adequate equivalent. Some of the English words should be of such a nature as to confuse the poorly prepared student. Examples are:

gladius	(1) glad	(2) flower	(3) wound	(4) sword	(5) spear
hora	(1) hoary	(2) hour	(3) horse	(4) night	(5) hair

There is room for considerable ingenuity in assembling good "confusion" words or "distractors." An excellent source for wrong responses is the actual pupil errors in written work, or in "recall" types of vocabulary tests. It should be noted also that students who do not *know* the correct response tend to look for and to select a word whose spelling suggests that of the Latin word, or a synonym of such a word. In the first illustration above, for example, the word "glad" is likely to prove a good foil for this reason, as might also the word "happy." "Flower" is also likely to attract students because

¹¹ Catherine M. Haage, *Tests of Functional Latin for Secondary School Use*. Doctoral thesis of the University of Pennsylvania. Philadelphia, 1932. Pp. 192. See p. 162, Table xxix.

"gladius" suggests "gladiolus." By careful and ingenious selection of the "distractors," the student who guesses can be almost invariably misled, as he should be, and the number of correct guesses can be reduced far below that which would result by pure chance.

In such a multiple-choice technique, the English equivalent is to be underscored, or (preferably) its number may be written in a blank provided at the right or at the left of the page.

The vocabulary test recently developed by Miss Haage¹² requires the pupil to supply a word missing in a Latin sentence by choosing from five that are suggested. An example is:

Gladio et sagittis patriam meam —

(1) prohibeo (2) clamo (3) defendo (4) doceo (5) capio

It is clear that this is a test of comprehension as well as of vocabulary. It is to be noted, however, that in composing the sentences the author was reasonably careful, at any given level, to remain within the appropriate limits of the College Board word list. Similarly, in adopting either of the two techniques mentioned, a teacher should remain within the limits of the class experience, unless he may wish to combine with a vocabulary test proper, a test of the pupils' ability to make inferences about English derivatives from Latin words which are unfamiliar to them. But such a test would preferably have its own identity.

A third type of vocabulary test familiar to all teachers is the "recall" type, in which the pupil gives an English equivalent for a Latin word, or vice versa. About the same number of items can be given in the same length of time if this technique is followed, but the scoring cannot be so rapidly and accurately done. Pupils should be able in 15 minutes to complete a

¹² *Ibid.*, pp. 100-31.

fifty-item vocabulary test, whether constructed according to the first or the second procedure suggested above. However, the teacher will find it easier to follow the first, for the second procedure involves either finding or inventing suitable sentences within the vocabulary limits, and also choosing possible Latin equivalents which fall within these limits and which, at the same time, offer to the student something like a real problem in making his selection.

No matter what procedure is followed in making the examination, whether the model set by the examining agencies mentioned above, or certain of the other patterns of the so-called new-type tests, it is clear that a specific test of knowledge of vocabulary and of idiom is highly useful. Brief examinations of this sort, administered fortnightly or monthly, have a diagnostic value and aid the pupil in recognizing his deficiencies. A longer vocabulary test at the end of a semester reveals progress in a fundamental element of the course and also, as has been pointed out, helps indirectly to measure progress in reading ability.

2. Grammar

Knowledge and understanding of grammatical phenomena have traditionally been tested by:

1. Requiring a translation from Latin into English, which involves application of grammatical knowledge, as well as of vocabulary and idiom, and that elusive power, both in Latin and in English, which we call "speech-feeling." The pupil who possesses this speech-feeling in the two languages is guided more surely by "a succession . . . of anticipations and fulfillments," as the Report of the Classical Investigation puts it, to a comprehension of what the Latin writer has said, and then, in turn, to selecting a means of expressing this in English.

2. Asking questions about words found in the passage to be

translated; their forms ("decline," "conjugate"), or their uses, or both.

3. Requiring a translation from English into Latin, which involves recall of vocabulary and idiom, recognition of the need for specific forms and recall of those forms, and some feeling for Latin word order.

4. Calling for the inflections of given words — "decline," "conjugate," "compare," etc.

The measurement of grammatical knowledge and understanding by the first of these devices is open to the double objection that the grammatical element is inextricably bound up with the other elements, and that uniform scoring is well-nigh impossible. The third device, translation into Latin, presents in differing degrees the same difficulties. The other two procedures, if the items involved are significant in the course, are liable to few criticisms from the standpoint of validity, but are less easily scored than other techniques.

Examiners probably do not make sufficient allowance, particularly during the "elementary" or two-year stage, for the fact that pupils are none too sure in handling grammatical terminology. For example, when asked to conjugate *transibant* in the present subjunctive active, or *summovebant* in the future indicative passive, or to name and give the reason for the mood of *vinceret*, or to decline the plural of the comparative of *multi* in all genders, many pupils may be unable to link the correct forms with the names of these forms. Emphasizing the functional rather than the formal, Miss Haage, in the study referred to, avoided all terminology in her "forms" test. She framed sentences in which the function of an underlined word (noun, pronoun, adjective, adverb) or phrase is to be indicated by answering such a question as "Whom or what does someone see, say, take, etc.?" A specimen sentence:

Procul *villam* pulchram videtis.

She lists ten questions involving case relations — “from whom? what? for whose benefit or to whom? by whom? etc.” — and a last item with the rubric “no question.” Each question is designated by a letter or number, and each underlined item in the 25 sentences is to be catalogued A, B, C, (or 1, 2, 3) if it illustrates an answer to one of the ten questions, or by K (or the number 11) if this construction fits none of the ten questions. The scoring is completely objective and the teacher may increase or diminish both the number of the questions, thus involving more or fewer usages, and the number of sentences containing phenomena to be analyzed.

A considerable number of techniques for testing on forms and syntax are illustrated in Ullman and Smalley's *New Progress Tests in Latin*,¹³ as follows: supplying the ending for an indicated form of a noun, pronoun, verb, etc.; choosing, out of four or five English versions, the correct one to fit a Latin verb form; giving the proper Latin form of an italicized English word in an English sentence (“Where is my *brother's* book?”) without translating the sentence; choosing from four or five Latin forms the one that suits a given situation (“Where is the *large* island? 1. magna, 2 magnam, 3 magnas, 4 magnae”),¹⁴ and the like. Such exercises include some testing of usage as well as of forms, but it is hardly essential to divorce the two rigidly.

A syntax matching test may be constructed on the pattern of the Godsey Latin Composition Test.¹⁵ A number of “rules” are stated, as “The direct object of a verb is in the accusative case,” and a group of Latin sentences is given in each of which an underscored word or phrase is to be referred by a

¹³ New York: The Macmillan Co., 1934. Pp. 122.

¹⁴ This technique is exemplified in the 100-item grammar test in the *Cooperative Latin Test*, Form 1933, prepared by the Cooperative Test Service.

¹⁵ Edith R. Godsey. Yonkers, New York: World Book Co., 1926. See also Ullman and Smalley, *op. cit.*, pp. 87-88.

number or a letter to the appropriate rule. The technique of the Haage "forms" test already mentioned is a variant of this one.

Examinations on forms and syntax constructed according to one or more of such techniques are more interesting to students than the ordinary type, and at the same time may be scored uniformly and accurately. Moreover, since the forms and the syntactic phenomena appropriate for each semester of the course have been generally agreed upon,¹⁶ the teacher can be sure of the ground to be covered in a given semester examination.

3. Reading

The pronouncement of the Committee of Twelve of the American Philological Association, that the student should be carefully trained to take in the meaning of the sentence in the order in which it stands, before translating it, has already been referred to. The *Report of the Classical Investigation*,¹⁷ as has been noted, endorses this principle and sets forth at length the procedures by which pupils may be trained to approach a Latin sentence in the right way. We have also observed that despite these utterances, and despite the general approval of the principle by Latin teachers, it is usually ignored in classrooms and in the preparation of examinations, whether by individual teachers or by such agencies as the College Entrance Examination Board and the Secondary Education Board. This is not the place to discuss the principle involved, but surely a translation test, involving as it does so many other elements, cannot be considered a measure of the ability to read Latin directly. Investigators who have attempted to

¹⁶ See *Report of the Classical Investigation*, pp. 157-62; *Study of the Secondary Curriculum*, the Secondary Education Board, pp. 87-90; *High School Curriculum Reorganization* (North Central Association), pp. 152-53, 159-61, 165, 169-70.

¹⁷ Pages 188-97.

construct tests of reading ability which at least do not directly contravene the principle of teaching pupils to read Latin as Latin, have utilized particularly the paragraph-question type¹⁸ and the multiple-choice technique. There is, of course, no guarantee that pupils do not first translate the passages, particularly if they have been expected in their daily work to show by translation the extent of their preparation. It is true, however, that the expectation of a translation test of their reading ability will encourage them, and very naturally, to consider that the only proper way to read a Latin passage is by translating it into English. Consequently, if the aim of reading Latin as Latin is desirable and attainable, progress in doing so should be measured by other than a translation procedure. The "deciphering" behavior, in reading Latin, of the pupils whose eye-movements and comprehension were studied by Messrs. Judd and Buswell¹⁹ was due in large part, no doubt, to the fact that the pupils had received no training in any other than a "peek and poke" manner of preparing lessons, that they had always been required to give evidence of their preparation by translation in class, and knew that a similar type of reading examination awaited them at the end of each semester.

A second serious objection to testing reading ability through translation is the impossibility of scoring a number of translations in a uniform way. From a pupil's English version of a sentence or passage, it is frequently impossible to tell whether the original has not been understood or whether the trouble lies in the pupil's inability to express himself in English. Practice in translation for the sake of improvement in English has its values as a part of the pupil's experience, but

¹⁸ B. L. Ullman and T. J. Kirby. *Latin Comprehension Test*. Iowa City, Ia.: Bureau of Educational Research, State University of Iowa, 1922.

Catherine M. Haage, *op. cit.*, pp. 101-14.

¹⁹ C. H. Judd and G. T. Buswell. *Silent Reading: A Study of the Various Types*. Chicago: University of Chicago Press, 1922.

it is the rôle of an examination to measure achievement in the subject matter, to reveal as precisely as possible the student's growth in knowledge or in skills, or in both. Making a suitable translation of a passage entails a good deal more than just understanding what the passage says. If skill in *translation* as such is one of the aims, it is proper to exact a translation test, despite the difficulty of scoring it. If, however, our purpose is to test the pupils' ability to *read*, in the commonly accepted sense of the word, then a translation test alone is not enough.

One obvious reason for the vigorous survival of translation as a test of reading ability is, of course, the comparative ease with which such a test is prepared. Furthermore, teachers are so accustomed to giving grades which are mere approximations that they are not unduly shocked by their inability to differentiate with some precision between the performance of their pupils on examinations.

In addition to the paragraph-question technique, several other types of devices for testing comprehension may be enumerated:²⁰

1. A series of statements in Latin, each of which is to be checked in an indicated way according to whether it is true or false.

2. A Latin sentence in which a missing word or phrase is to be supplied from several that are suggested. By this device, emphasis may be thrown on comprehension or on vocabulary or on grammatical points, by the selection of the sentences and the suggested words or phrases employed.

3. A passage in Latin, followed by groups of statements, in English or in Latin, about its content. The pupil is to select in each group the statement or statements which agree most

²⁰ Cf. Ullman and Smalley, *op. cit.*, pp. 107 *et seq.*

closely with the passage. This is a variant of the true-false technique.

4. A Latin sentence, and a translation of it which contains an error. The student is to correct the error. Or several versions may be proposed, the student to select the best. The latter procedure enables a student to express his judgment on a larger number of sentences in a given time, and enables the teacher to score the results more objectively.

4. Derivatives and Word Study

The belief that the study of Latin contributes, or may be made to contribute directly to a better knowledge of the English language is widespread and is duly emphasized in the *Report of the Classical Investigation*,²¹ which calls to the attention of teachers the value of specific efforts in this direction. The report recommends strongly "that tests and examinations should regularly include questions on the more important aims of the subject," and points out that fewer than half of the examination papers assembled in connection with the investigation contained questions bearing on the objective under discussion. Such questions appear in the examinations set by the College Board only for the more advanced years, and there are none in available question papers of the Secondary Education Board and of the New York Regents.²² Again it must be insisted either that progress in this direction must be measured as a part of the whole picture of student attainment, or that statements about its importance should no longer figure prominently in formulations of objectives.

²¹ Pages 42-44, 210-17. Interesting and useful material is contained in the Latin section of *High School Curriculum Reorganization*, *passim*. See also G. M. Ruch and G. A. Rice, *Specimen Objective Examinations*. Chicago: Scott, Foresman & Co., 1930, pp. 248-59 (derivatives, spelling, vocabulary, grammar, comprehension).

²² But see above, note 9. This element is present in the Regents' examination for 1934.

While taking this position, we must not go to the other extreme and make class exercises in Latin too largely a study of derivatives, word-formation, and the like. Excessive zeal for the cause may prove as fatal as virtual neglect. But the recommendation of the College Board that about 450 Latin words should be acquired by the pupils during each year of the school course²³ serves as a useful check on any tendency towards overemphasis.

The procedures advocated by most teachers for learning vocabulary are closely related to the topic under discussion. These are:²⁴

Associating a new Latin word with English derivatives or with related Latin words before the word is met in a sentence.

Determining the meaning of a new Latin word from context, association with English derivatives, or association with related Latin words as the new word is met in a sentence.

As a corollary of the above, the Report lists more specifically certain ways in which class exercises may contribute to the end in view:

Encouraging pupils to discover independently new derivatives from Latin words already learned.

Encouraging pupils to discover in their English reading derivatives discussed in class.

Encouraging pupils to discover independently new derivatives from Latin words specially assigned.

Encouraging pupils to use in sentences derivatives discussed in class.

Definite assignment of English derivatives for explanation on the basis of their etymology.

Each teacher should keep a record of the specific cases in which one or more of these procedures has been applied, so

²³ See also *Report of the Classical Investigation*, p. 209.

²⁴ *Report of the Classical Investigation*, pp. 207, 209, 210-13.

EXAMINATIONS IN THE FOREIGN LANGUAGES

that when examination time comes he can draw on a list of items for this portion of the test. He may utilize in the examination either a specific item that has been treated in class, or an item new in form but closely analogous to one that is familiar. Useful testing devices are:²⁵

1. An English sentence containing a derivative, the meaning of which is to be indicated by a multiple-choice technique.

Sample: Your memory is more *tenacious* than mine.

1 stubborn, 2 retentive, 3 firm, 4 receptive, 5 yielding

2. A list of English words opposite each of which the pupil is to write a closely related Latin word.

Sample: regal *rex*

3. A proportion the missing member of which is to be supplied.

Sample: victory: victoria :: *perfidy*: perfidia

4. A number of Latin words which are to be changed so as to make them into English words.

Samples: $\begin{matrix} & e & & ce \\ f\bar{a}m\bar{a} & & & d\bar{i}l\bar{i}g\bar{e}n\bar{t} \end{matrix}$

5. A list of Latin words, and a list of the corresponding English derivatives but in a different order. The pupil is to rearrange one list so as to bring together each Latin word and its derivative.

6. A list of Latin words such as *accēdō* and such English words as *announce* for which the original form of the prefix is to be supplied.

7. A list of English words, the meaning of each of which is to be given so as to show its derivation.

Sample: *impervious*, not allowing a way through.

²⁵ See Ullman and Smalley, *op. cit.*, pp. 89-106.

8. A list of compound Latin words and English words, for each of which the pupil must give a simple Latin verb:

Sample: *inscriptus* — *scribō* concession — *cēdō*

9. The formation of Latin compound verbs from parts given.

Sample: *ad* — *cēdō* *accēdō*

5. The Historical-Cultural Element

Whether we designate the material included under this heading as Roman civilization or culture or *Realien* is of small importance. We mean by it such material bearing on various aspects of ancient Rome — geography, history, religion, social organization, ways of living, public and private life — as it is appropriate to present to pupils of secondary school age. The medium of presentation may be the class textbook, especially the reading material, with suitable illustrations, notes and commentary, and readings in English.²⁶

Here again each teacher should keep a record of the material of this kind, whether informational or interpretative, that is presented to a given class during a given period, and should draw from these records the items to be included in an examination. And here again, as in the preceding section, a warning against undue zeal is pertinent. One cannot expect this aim of the course to be achieved unless some specific provision is made to that end, but it should receive a suitable share of the teacher's attention, and no more. The enumeration furnished by the *Report of the Classical Investigation* (pp. 152-56) may profitably be utilized by teachers, each one allocating the items to that part of a given course of study to which they belong, as determined by the textbook, the reading matter, the teacher's commentary, and the like.

In constructing examinations on this aspect of the course,

²⁶ *Report of the Classical Investigation*, pp. 151-56, 204-06. See also for useful material the Report of the Secondary Education Board and *High School Curriculum Reorganization*.

EXAMINATIONS IN THE FOREIGN LANGUAGES

the teacher may make use of devices similar to those already suggested in other connections:²⁷

1. Statements to be checked as true or false:

Juno was the daughter of Jupiter.

2. Statements to be completed:

The chief Roman god was ———.

3. Groups of statements about government, the city of Rome, amusements, etc., to be checked according as they are false or true. A variant of number 1, the items being grouped about different topics.

4. A list of proper names, accompanied by a list of attributes or explanations or judgments arranged in a different order from the list of names. The pupil is to match those that belong together, whether by actually writing the lists or by entering a number in an indicated place. For example:

- | | |
|------------|-----------------------------|
| 1. Mercury | (3) God of war |
| 2. Diana | () Goddess of love |
| 3. Mars | () King of the lower world |

5. A variant of number 4, consisting of a series of statements followed by a list of names, each of which is to be matched with the proper statement.

CONCLUSION

It will be apparent to every reader that much emphasis throughout this presentation has been placed on two points. The first and more important of these is that examinations should reflect as completely as possible the aims of the course. If there is no way of testing progress toward a stated goal, it would appear to be wiser to revise the statement of aims.

²⁷ See Ullman and Smalley, *op. cit.*, pp. 113-22.

Either the objective in question should disappear from the list, or it should be placed among the imponderable values, such as the development of desirable attitudes, habits, and the like, which may be legitimately hoped for but may not be measured by any device at the command of most members of the profession. And surely the amount of progress made toward the objectives enumerated above can be measured on a comparative basis. The second emphasized point is the great advantage of so constructing examinations that they can be scored in a uniform way, thus making it possible really to compare the score of one pupil with that of another. A third point should be added to these two: namely, that once or twice during the school year an examination so constructed should be administered to *all* pupils, from the first-semester level up. The distribution of the resulting scores would prove rather surprising in many cases, and would illustrate so vividly the overlapping between classes that the teacher and the school authorities might feel almost obliged to take some action, whether to promote or to put back certain pupils, or perhaps even whole classes; to give remedial instruction on particular topics; or to grant special privileges to high ranking pupils while the less well-advanced are being trained specifically to remedy their weaknesses.

The use of standardized tests²⁸ from time to time is also commendable, for it enables the teacher to draw very precise comparisons between the scores of his pupils and those of a rather large number of others. The lack of a sufficient number of "forms" or equivalent versions of available tests would alone prevent the frequent administration of tests of this kind to the same pupils. But even if many forms of the existing

²⁸ A list of Latin achievement tests, prepared by Professor V. A. C. Henmon, is given on pages 490-91, as a convenience to teachers who seek to inform themselves further about testing techniques.

tests were available, teachers would still wish to test their pupils on a particular section or unit of their own course, and for this the standardized tests are naturally not suitable.

All being considered, most of the testing, the country over, will for a long time be done by means of "home-made" examinations. Those who follow the recommendations of the Classical Committee in regard to course content in respect to grammar, reading, and cultural-historical material, and the College Board list in respect to vocabulary, are in an excellent position to profit by the foregoing proposals for constructing examinations. The minimum basic subject matter is determined from semester to semester, and teachers have only to devise the specific items which will elicit from their pupils the appropriate response. Furthermore, such examinations, if filed away with an adequate record of the results of administrations, may be compared with other examinations constructed so as to be closely equivalent, and the school will thus be in possession of roughly standardized tests and test results, which will do more to minimize fluctuations in class standards from semester to semester, from year to year, and from teacher to teacher than almost any other readily available means. This alone would be a genuine accomplishment, for all inquiries into the subject, both in Latin and in modern languages, have shown the great disparities in attainment, in the same school and in different schools, by pupils and by classes ostensibly on the same level of advancement. It is clear, furthermore, that these disparities cannot even be detected unless more refined measuring instruments are used than the conventional type of examination. It is true that the readers of the College Board examinations do, to a considerable extent, triumph over the inadequacies of the examinations and do achieve a reasonable degree of uniformity in their scoring, but they succeed in doing this despite the techniques followed

in the examinations rather than because of them, and few teachers are in a position to emulate this group of readers.

QUESTIONS FOR DISCUSSION

1. Examine critically the recent C.E.E.B. examinations in Latin. Compare your analysis with that given in the text (pp. 266-67) and point out the resemblances and the differences.
2. Study in the same way the recent Latin examinations set by the New York Regents.
3. What specific aims for instruction in Latin do you judge that the makers of these examinations had in mind?
4. Compare with the College Board word list (cf. S. M. Hurlbut and B. M. Allen, *A Latin Vocabulary for First and Second Years, A Latin Vocabulary for Third and Fourth Years*. New York: American Book Co.) the vocabulary (Latin-English and English-Latin) needed by the pupil in writing these examinations. (Questions 2 and 3). To what conclusion do you come?
5. With a familiar first-year or second-year textbook as a basis, make two 20-item vocabulary tests, using the two techniques exemplified on pp. 271-72.
6. Secure from a neighboring school a written translation made by a class of one or more representative passages in Latin. Arrange with several of your classmates that each of you will score the translation independently. Compare the scores thus assigned and try to account for the divergences. It is better to have copies typed for distribution to the participants.
7. Follow the same procedure with a group of sentences which pupils have translated from English into Latin.
8. Make out two or more examinations in grammar in which you utilize the four techniques exemplified in Ullman and Smalley's *Progress Tests* (p. 275).
9. Similarly, make a ten-item test following the Godsey Composition test technique (p. 275).
10. Using a brief Latin passage, prepare a test in which you use the "best statement" technique for testing understanding of the passage (p. 278).

EXAMINATIONS IN THE FOREIGN LANGUAGES

11. Taking four or five Latin sentences, exemplify the "missing word" technique (p. 278).
12. Make a list of all the Latin words in a given number of lessons (5 to 10) of a first or a second year textbook, and list the English derivative or derivatives of each, where possible. Select from these two lists the items that you think it desirable and practicable to teach.
13. With these lists as a basis, make nine 5-item tests, each of which illustrates one of the techniques exemplified on p. 281.²⁹
14. Take a first or a second-year textbook and compile a list of all items of an historical or cultural nature contained in any segment of fifty pages of the book.
15. On the basis of this list, make a fifteen-item test in which each of the five techniques on p. 283 is illustrated.
16. Secure a number of specimens of semester Latin examinations from neighboring schools. In the light of the preceding discussion, evaluate the content and the techniques used.
17. Using the same subject matter, attempt to rebuild two or three of these examinations, according to the techniques suggested above.
18. Secure copies of a number of the tests listed on pp. 490-91, preferably those on which at least tentative norms are available. Administer them if possible in one or more high-school classes and study the results.
19. Look up in the *Classical Journal* and other likely places the articles on testing that have appeared since 1925, and summarize them.
20. Why are tests that are scored objectively so rarely used in classroom examinations?

II. MODERN LANGUAGES

From the preceding chapters it is evident that an examiner must have clearly in mind the aims which pupils are expected

²⁹ For a 35-item test illustrating the first type, see Ruch and Rice, *Specimen Objective Examinations*. Chicago, Scott, Foresman and Co., 1930. Pp. 249-51.

to attain in his subject, or at least in that portion of the subject on which a test is being set at a particular time. It is also clear, or should be before this chapter is concluded, that measurement of progress in that extremely complex process which we call learning a foreign language involves more factors than modern language teachers have usually assumed, that it demands more planning and more ingenuity and, consequently, more time and effort on the part of the examiner.

It has been pointed out by advocates of changes in teaching, especially in countries where examinations play a larger rôle than in American schools, that to bring about reforms in teaching a subject, one must begin by reforming the examination procedures. That, in a sense, is the thesis of this volume, but in the discussion of foreign language examinations which follows, the committee has endeavored to avoid controversial issues, except in so far as these issues are inherent in any invitation to teachers to consider critically their aims and to measure with some accuracy the degree to which these aims are attained. When, for example, one urges teachers to measure the attainment of their pupils in ability to *read* rather than in ability to *translate*, one raises by implication the whole question of what reading is and of the procedures which are most favorable to its development. However, having once taken cognizance of this fact, this section will be limited to its immediate purpose.

OBJECTIVES

The aims, or desired results, of modern language teaching have been formulated by innumerable individuals and committees, from the *Report of the Committee of Twelve* in 1898 to the present day. A tabulation made in 1926 from modern language bulletins issued by twenty-two states yielded the

EXAMINATIONS IN THE FOREIGN LANGUAGES

following list of the most frequently occurring items, in the order given:

1. Ability to read.
2. Ability to write.
3. Ability to speak.
4. Acquaintance with the history, the literature, the people of the foreign country.
5. Ability to understand the foreign language when spoken.
6. Mastery of the grammar of the foreign language.
7. Ability to translate from English into the foreign language.
8. Better understanding and appreciation of the English language.
9. Mental discipline.¹

In 1928 the Committee on Investigation of the Modern Foreign Language Study formulated immediate and ultimate objectives for the first two years of study, corresponding to the "elementary" course of the Committee of Twelve and the College Entrance Examination Board, and also for courses of longer duration.² For a two-year course, the statement of immediate aims is as follows:

Progressive development:

1. Of the ability to read books, newspapers, and magazines in the modern language within the scope of the student's interest and intellectual powers.
2. Of such knowledge of the grammar of the language as is demonstrated to be necessary for reading with comprehension.

¹ From an investigation by Sturgis Leavitt and C. A. Stoudemire, utilized in Algernon Coleman, *Teaching of Modern Foreign Languages in the United States*. Publications of the American and Canadian Committees on Modern Languages. New York: The Macmillan Co., 1929, p. 8.

² See Coleman, *op. cit.*, pp. 107-08.

3. Of the ability to pronounce correctly, to understand (aurally), and to use the language orally within the limits of class materials.

4. Of a knowledge of the foreign country, past and present, and of a special interest in the life and characteristics of its people.

5. Of increased knowledge of the derivation and meanings of English words, of the principles and leading facts of English grammar, and of the relationships between the foreign language and English.

For the more advanced group, the immediate aims as listed in the Committee's statement assume further progress in all the abilities named above, and, in addition, such progress in ability to write the language as to make of this an additional and specific objective. It is evident that the Committee envisaged gradual progress on the part of pupils throughout the period of study in ability to read the foreign language, to understand it when spoken, and to make themselves understood by others when speaking and when writing it — progress which would bring their attainment in the foreign language nearer to their attainment in the mother tongue, although at no time would attainment in the foreign language equal attainment in the vernacular, except, possibly, in ability to read.

Since 1929, there have been at least four significant attempts at a restatement of objectives.³ The resulting statements, while they differ from one another in important respects, have

³ Alice Corell and others. *Tentative Syllabus in Modern Foreign Languages*. Albany, New York: University of the State of New York, 1931, p. 138.

Lawrence A. Wilkins and others. *Syllabus of Minima in Modern Foreign Languages*. New York: Board of Education, 1931, p. 146. Howard T. Smith and others. *Report of a Study of the Secondary Curriculum*, Milton, Massachusetts: The Secondary Education Board, 1932, pp. 100-63.

L. A. Webb and others. *High School Curriculum Reorganization*. Ann Arbor, Michigan: The North Central Association of Secondary Schools and Colleges, 1933, pp. 175-214.

much in common with each other and with those contained in the reports referred to above. Two of them, those issued by New York State and by the Secondary Education Board, retain the disciplinary principle enunciated by the Committee of Twelve. They also consider the first two years as providing primarily a foundation period for additional study of the subject in school or in college. The syllabus for New York City places reading ability frankly in the foreground, does not propose ability to write the language as an aim in itself, but does propose as aims the attainment of "a reasonably fluent and accurate pronunciation," "an introductory knowledge of the foreign country," and "ability to grasp readily thought expressed in the foreign language . . . in speech."

It is evident that, in varying degrees and by varying procedures, most modern language teachers propose to have their pupils learn to read texts in the foreign language; to make themselves intelligible when pronouncing the language; to understand the language when spoken; to use it orally and in writing — within narrow limits, to be sure, but in accordance with current usage; to learn more about the foreign people and the foreign country than if they were not students of the language; and, at the same time, to make more progress in their control over the mother tongue than non-foreign-language students would do.

This is not the place to debate the soundness of the individual items of this list of aims, or to point out the qualifications which a wholly realistic view of the problem might suggest at various points. Our task is merely to enumerate the objectives which most members of the profession accept, and, on the basis of this enumeration, to offer a program for testing the degree to which these objectives are attained.

The attainment of the aims listed above implies, in varying degrees, the following: learning the sounds of the language —

EXAMINATIONS IN MAJOR SUBJECT FIELDS

to make them and to recognize them when heard; learning the requisite words and idiomatic expressions; acquiring familiarity with the way in which foreign writers and speakers put their words together, in respect both to the forms and the arrangement; relating the foreign vocabulary to the mother tongue and, if appropriate, to another foreign language, both in form and in meaning; reading and hearing about the foreign people and their country, past and present, and reading books that are representative of the foreign people in some significant, although not necessarily profound way.

EXAMINATIONS

Current Practices

If the foregoing statement of aims, and of what students must do in order to attain them, is representative, it follows that we have taken the first step in determining the nature of a testing program, for, as the preceding chapters have made clear, examinations should aid teachers in ascertaining the progress that is being made toward the attainment of objectives. Or, to express the same truth in a different way, they should measure progress in the knowledge of subject matter and in the attainment of the skills which it is the aim of the teacher progressively to develop. It is generally admitted that prevalent types of examinations in modern languages fail to accomplish this, partly because of difficulties inherent in the situation, and partly because teachers have not, in general, analyzed with sufficient care the purposes and the shortcomings of their examinations and have not given sufficient thought to ways of remedying these shortcomings.

Two recent studies,⁴ based on wide administration of care-

⁴ Ben D. Wood, *The New York Experiments with New-Type Modern Language Tests*, and V. A. C. Henmon, *Achievement Tests in the Modern Foreign Languages*.

fully standardized tests, have shown in great detail that pupils of widely differing attainment in the subject are enrolled in the same classes, and that corresponding classes in the same school and in different schools vary greatly in their knowledge of the subject matter. "Finally," says one writer after surveying the evidence, "we may reiterate that at present nothing less than chaos prevails in the classification of our modern language students. The fact that a course is of a given length may have and usually does have little relation to the knowledge of the subject attained by a given class. Many two-year classes are superior in actual attainment to many three-year classes."⁵ The variations among individuals are, of course, even more numerous and striking, with the result that the fact that an individual has "had" two years or three years of a modern language in school or in college provides little or no evidence in regard to attainment in the subject. Consequently, the widespread assumption on the part of many colleges which admit on certificate that a two-year high-school course fulfills the foreign language requirement for admission, or for graduation, is, in many cases, based on fiction.

One well-known college formulates thus its foreign language requirement for exit from the junior college stratum: an examination which demands "the mastery of a foreign language at the level of attainment expected of a student who offers two acceptable entrance units in a foreign language, unless the student has offered two acceptable entrance units in a foreign language." The adjective "acceptable" in this statement may or may not be significant. Its value depends on the means Publications of the American and Canadian Committees on Modern Languages. New York: The Macmillan Co., vol. I, 1927, and vol. V, 1929. See also Coleman, *op. cit.*, chapter III.

⁵ Coleman, *op. cit.*, p. 231. For corroborating data see Wood, *op. cit.*, pp. 155-67, 173, 178, 183-97; Henmon, *op. cit.*, pp. 131-32, 145-46, 176-207.

used by the institution to give it significance. Another institution recently announced that a placement test will be administered as a substitute for, or to give meaning to, the conventional two-year requirement. The University of Wisconsin recently adopted a policy whereby tests for measuring attainment in modern languages may be substituted for the time requirement.⁶ Two levels are proposed: "proficiency" or advanced knowledge, and "reading knowledge." The latter is understood to mean "the ability to pronounce the modern language and to interpret adequately modern prose of average difficulty." The former level is attained when the student can give evidence of "(a) adequate comprehension of representative passages from classic and modern authors, which may include matter taken from the student's major field, (b) the ability to understand and pronounce simple phrases in the spoken language, (c) some knowledge of the history of the literature and culture of the foreign people." Apparently, these two attempts to be more explicit in the statement of achievement are rather isolated instances, but they suggest growing dissatisfaction with the practice of measuring attainment in the subject by the time spent in class.

What are the chief defects in prevalent testing procedures? Perhaps the reply to this question may be expressed in five statements:

1. Few or no examinations or groups of examinations test progress in all the types of knowledge and of skills which are embraced by the aims enumerated above.
2. In few cases are comprehensive examinations, based on an adequate sampling of the subject matter, given to different

⁶ C. D. Zdanowicz, *Modern Language Journal*, xv, 5 (February, 1931), 354-59; F. D. Cheydleur, *French Review*, vol. vi, no. 3 (February, 1933), pp. 190-214, and no. 4 (March, 1933), pp. 282-300. The quotations are taken from Cheydleur's article.

semester- or year-groups at the same time so as to measure the achievement of all with a common instrument.

3. Rarely are tests so constructed as to measure with relative accuracy attainment at the different semester- or year-stages.

4. Because of a natural tendency to follow beaten paths, and because of the considerable difficulties encountered in breaking new ground, teachers do not often set examinations which can be so scored as to reflect the relative importance of the different aims which they profess.

5. Tests rarely are so devised as to be scored with a high degree of uniformity and accuracy.

Let us consider these statements in reverse order. We have no reason to believe that the last statement can be disputed. The evidence given by Wood and by Henmon, to mention only two studies of the subject, suffices to convince the most obdurate. With respect to the fourth statement, it is clear that the normal examination gives little weight to reading and to oral attainment, even in classes where these two aims are professedly the most important. As regards the third statement, our inability to posit with some definiteness the kinds and the degree of knowledge and of skill which represent fourth-semester or fifth- or sixth-semester attainment is too well known to need further comment.

The truth of the second statement, that uniform and comprehensive examinations are rarely given, is so much a matter of common knowledge as to call for no proof. Indeed, most teachers protest against the administration of such examinations. They are wedded to the assumption that differences in the length of the period of study correspond closely to differences in attainment, and are opposed to allowing their students to try their mettle on any question which involves material that they have not specifically "had." A complete

examination program would naturally include both progress or diagnostic tests, of which the items are selected from the materials which the pupils have "had"; and achievement tests, made up of material ranging from simple to difficult and covering the whole span of the high-school program or of the first two years in college.

The truth of the first of the five foregoing statements is almost equally a matter of common knowledge. To be sure, the absence in most testing programs of specific measures of oral and aural skills is explicable in large part by the difficulty of constructing tests for this purpose. But even in the case of students whom we meet every day, whose oral and aural skills we test constantly, we rarely give any systematic weight to this element when assigning a grade. Professor Ford⁷ found that in the Montreal schools, where oral proficiency is considered important and is, theoretically, graded separately, "grammatical knowledge is clearly the most important factor in determining the oral mark."

Furthermore, few examinations contain a representative test of vocabulary as such. Almost every item of conventional examinations makes some contribution in this direction, but these provide usually only casual samplings of narrow scope. Most teachers of modern languages will agree that poverty of vocabulary is probably *the* major handicap of our students, whether for reading or for oral and written expression, and that the principle of "usefulness" as a guide to the choice of vocabulary, which first came to public attention through investigations by educators of ways of accelerating progress in learning to read and to write English, is of capital importance. A number of the new textbooks in the modern

⁷ *Modern Language Instruction in Canada*, Part II, Publications of the American and Canadian Committees on Modern Languages, vol. vi. Toronto: University of Toronto Press, 1928, pp. 837-46.

language field offer evidence in support of this statement and provide means for a more systematic application of the principle,⁸ which, however, is rarely reflected in examination papers.

It is also true, paradoxical as the statement may seem, that few modern language examinations test reading ability, except in an extremely narrow sense of the word. A passage, chosen somewhat casually, to be translated into English is a familiar device. If this passage happens really to represent in vocabulary and idiom, in form and in content, the work of the class, it has testing value; but how often is the passage chosen subjected to an analysis on the basis of such criteria? In some cases, teachers propound questions on the content of a passage instead of requiring a translation. Such a procedure has the advantage of putting the student more nearly in the normal attitude of a reader, and permits also of setting eight or ten passages, each with appropriate questions to be answered in the mother tongue. This alone is of value, providing, as it does, a more nearly adequate sampling of the student's reading power with different types of texts which may present a fairly wide range of vocabulary, idiom, and grammatical phenomena. Such a technique is not often used, partly because teachers doubt its efficacy to measure reading ability, partly because suitable passages are not readily found, and partly, no doubt, because it implies equipment for mimeographing the examination.

It must be noted, however, that even by such a group of passages — understanding of which is tested by questions, as in the Thorndike-McCall paragraph-question reading test in English, or by a multiple-choice device, or by a "best-

⁸ See Algernon Coleman, *Experiments and Studies in Modern Language Teaching*. Chicago: University of Chicago Press, 1934, pp. 50-88. Current announcements by publishers of textbooks testify to the rapid spread of this idea.

answer" device — the learner's ability to read a story or a book continuously is not tested, nor does the procedure test the rate at which he reads. And surely these two factors are highly important. Even those who maintain that slow, intensive study of brief assignments is the only suitable type of reading practice at the early stages will admit that the desired goal is the ability to read appropriate texts consecutively and at a rate which approaches reading performance in the vernacular. It follows, then, that some attempt at measuring progress in this direction should enter into the testing program.

Proceeding further in our analysis, we observe that the conventional examination tests progress neither in knowledge of the foreign country and its life nor in general growth in linguistic knowledge and understanding as evidenced especially by relatively greater progress in English. The studies of Werner and of Woody⁹ tend strongly to discredit the optimistic claims made by modern language teachers with respect to progress toward these objectives. Indeed, we are accustomed to exercise faith rather than knowledge in estimating the progress made by our pupils, both on the cultural and on the linguistic side.

If, then, the conventional examination does not measure on a comparative basis oral skill, or aural attainment, or progress in the acquisition of vocabulary, or in reading ability in any broad sense of the term, what does it test? All teachers know the answer to this question. It tests chiefly knowledge of grammatical forms and usages. And as our class exercises are chiefly concerned with grammatical forms and usages, on the hypothesis that knowledge of these represents power to write and to speak the language, we have developed a consider-

⁹ In *Studies in Modern Language Teaching*. Publications of the American and Canadian Committees on Modern Languages, vol. xvii. New York: The Macmillan Co., 1930, pp. 99-184. See also Coleman, *The Teaching of Modern Languages in the United States*, pp. 95-98.

EXAMINATIONS IN THE FOREIGN LANGUAGES

able variety of useful devices, both for giving our students practice in grammatical forms and usages and for testing their success in learning them.

The papers set by the College Entrance Examination Board are representative. The June, 1931, examination in French, Cp. 2 (Two-year), is made up as follows:

1. A French passage of about 200 running words followed by nine questions in French, to be answered in French. (Three of these are linguistic.)
2. A French passage of about 200 running words to be translated into English.
3. An English passage of about 150 running words to be translated into French.
4. Ten cases in which the proper tense form of a given verb is to be inferred and used.
5. Five cases of missing pronouns to be supplied.
6. Ten cases of giving the opposite or the equivalent of words or of expressions.
7. Three questions involving the treatment of final consonants.

The corresponding German examination contains:

1. Two German passages to be translated into English.
2. One English passage to be translated into German.
3. One German passage followed by content questions to be answered in English.
4. Five sentences in which the proper auxiliary is to be placed, chosen from four suggested forms.
5. Six German words to be used in sentences invented by the candidate.
6. A question asking the student to tell, in German, the names of five interesting places in Germany and why they are interesting.

The corresponding test in Italian consists of:

1. An Italian passage to be translated into English.
2. Ten English sentences to be translated into Italian.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

3. Four tense forms of each of five verbs.
4. Nine Italian sentences to be revised, by supplying the proper finite verb form in four and the polite form of the pronouns in five.
5. Five substantives related to as many other words.
6. Five cases of syllable division.

The corresponding Spanish examination consists of:

1. A Spanish passage to be put into English.
2. A vocabulary test of 20 words.
3. Ten idioms to be translated into English.
4. Five English sentences to be translated into Spanish.
5. Twenty-five incomplete Spanish sentences to be completed by the proper version of given English words.
6. Ten incomplete Spanish sentences to be completed by suitable forms of given verbs.
7. Five cases of syllable division.

The June, 1933, examinations in the same subjects are of the same general type. It is evident that much ingenuity has been expended in constructing these examinations. There has been an effort to render certain portions more nearly objective in scoring, to broaden the nature of the reading test, and to reduce the rôle of grammatical forms and usages as such. In the German examination, one notes a slight recognition of the cultural aim, and the Spanish test contains a brief vocabulary section. Several of the examinations betray also an interest in pronunciation, perhaps as much as can well be conveyed in a written test. Nevertheless, this group of examinations has, on the whole, the defects noticed above, and even the praiseworthy attempt to provide a valid test of reading power in French is partly nullified by the provision that pupils must write their answers in the foreign language. Furthermore, much depends on the values given by the readers

to the different sections, and the printed copies give no hint of the weightings.¹⁰

A Suggested Program

The next step in treating the subject is an effort to propose a positive program,¹¹ with some illustrations. But it may be well, before doing this, to say a word about the rôle of the standardized test, of which a considerable number are now available to modern language teachers.¹² Useful as many of these are, they can hardly be expected to serve the teacher in all testing situations. Their diagnostic value in individual cases is probably not considerable, and teachers will continue for a long time

¹⁰ The C.E.E.B. examinations for 1934 contain, for French, German, and Italian, passages in the language followed by questions to be answered in English. In French, candidates at all levels were required to do this portion, a sort of embryonic uniform test. The Spanish examination, which adheres largely to translation, also contains a section required of all candidates: verb forms and phrases to be translated into Spanish (60 items), comprehension of definitions in Spanish (25 items), a type of vocabulary test (75 items). The Board's modern language examinations are, one would surmise, slowly evolving, although the form that they will eventually take is not clear.

The New York Regents' examinations for January, 1934, are distinctive in containing a "dictation and aural comprehension" section, which occupies one hour of the three and counts for 20 points. Questions on the passage for aural comprehension are to be answered in the foreign language, and credit is equally divided between comprehension of the passage and linguistic correctness. Each error in dictation counts one-half point, the word being the unit. As there are, for example, 82 words in the two-year German test, the scoring must present some difficulties. The rest of this examination consists of translation into English (20 points), forms (24 points), sentence completion (16 points). The examinations for three and for four years follow much the same pattern, except for requiring a 100-word composition in the foreign language on one of two or three topics.

The study by Crawford and Burnham referred to above (p. 268, note 9) raises the question of the predictive value of such examinations in modern languages.

¹¹ Useful suggestions and illustrations will be found in Florence M. Baker, *The Teaching of French*, chapter VIII. Boston: Houghton Mifflin Co., 1931.

¹² See the article by V. A. C. Henmon in Coleman, *Experiments and Studies in Modern Language Teaching*, pp. 191-218, and the catalogue prepared by the same author, for *idem*, "Analytical Bibliography of Modern Language Teaching," Chicago: University of Chicago Press, 1933, pp. 244-56.

to attempt to measure the success of their pupils in acquiring the specific skills or the knowledge appropriate to a given point in the course. Standardized tests are of particular value at the end of a semester or of a school year. Being provided with norms, they enable the teacher to institute comparisons between the scores of his pupils and those of a large number of others. If administered at the same time to all the pupils in a school or a system, they provide also a basis for ascertaining the amount of overlapping at the various levels and for reclassifying. Furthermore — and this is an important point — they can be of great assistance to the teacher in providing models of various testing techniques which he may apply in the construction of “home-made” tests, and, if only for this purpose, every teacher will find it profitable to make a collection of specimens of the standardized tests now on the market. Professor Henmon’s articles, which have been referred to above (note 12), should be consulted in this connection, since they present the most complete account of accessible standardized tests.

1. Oral and Aural Attainment

There are on the market three standardized tests of aural comprehension in French: American Council French Aural Comprehension Test, Forms A, B, C, D, E, by Agnes L. Rogers and Frances M. Clarke;¹³ French Test for Colleges: Aural Understanding, by Louise C. Seibert and Ben D. Wood;¹⁴ Lundeberg-Tharp Audition Test in French, Forms A and B, by Olav K. Lundeberg and James B. Tharp.¹⁵ But there are none in German and in Spanish.¹⁶ Consequently

¹³ New York: Bureau of Publications, Teachers College, Columbia University, 1932.

¹⁴ Yonkers-on-Hudson, N.Y.: World Book Company, 1928.

¹⁵ Columbus, O.: James B. Tharp, Ohio State University.

¹⁶ Messrs. Lundeberg and Tharp have standardized similar tests in German and in Spanish, but at this writing these have not been printed.

EXAMINATIONS IN THE FOREIGN LANGUAGES

teachers are virtually forced to adopt some means, however crude, of providing their own tests. The following suggestions may be helpful:

(a) Ask questions in the foreign language on well-known objects, places, persons, taking care to remain within the students' vocabulary limits. The replies are to be written in English.

Examples: Qu'est-ce qu'un cheval? Est-ce que Paris est en France ou en Allemagne? Qui a découvert l'Amérique?

(b) Display a large chart or picture before the class and make brief statements about it in the foreign language, which the pupils are to mark on their papers as being true or false.

(c) Define or describe familiar concepts in familiar language. The pupils are to show understanding by writing in English the name of the concept.

Examples: La capitale (*cr*, la plus grande ville) de la France. Celui qui a découvert l'Amérique. Ce qu'on porte sur la tête quand on va dans la rue.

These three procedures for testing aural comprehension lend themselves readily to objective scoring.

Another available technique for testing aural comprehension is to select a passage in the language of from 100 to 200 running words in length. Make sure that the vocabulary is within the limits of the pupils' experience. The passage should be narrative in character — a suitable anecdote would be best. Read this aloud to the class. Then have the class write their version of the story in English, or answer in English (written) two or more questions put orally in the foreign language and so framed that the answers will show whether the passage has been understood. Such a procedure should provide a rough score of aural comprehension. This test should be adminis-

tered on one or two days preceding the time allotted to the written examination.

To obtain an approximate score on attainment in pronunciation, have each member of the class, during a period of a week or so prior to the usual examination period, read aloud without interruption one or more passages similar to the above which call for some skill and some knowledge of the chief phonetic phenomena of the language. In French, for example, this would include elision, linking, phrasing, and such cases as are illustrated by: *plein, pleine; grand jardin, grand homme; deux, j'eus; ils sont, ils ont.*

In the two latter cases a score should be assigned to each pupil, and if the same passages are used at intervals of several months, some rough measure of progress should result. It may be objected that these scores would be, often, the merest approximation. The objection has point, but the probable error would be no greater than at present obtains in scoring the translation and free composition units of the normal examination, for the variations in scoring here are too well known to call for detailed proof. Two equally well qualified instructors, as has been repeatedly demonstrated, may vary as much as 20 to 40 points in grading a translation or a free composition unit; yet the value of such items in the examination scheme is usually taken for granted.

A procedure of this general sort, corrected by the instructor's daily contact with his students, may well be refined as the months go on, provided the instructor takes care to select or to compose suitable passages, to administer this portion of his test in exactly the same way to all his pupils, and to record his judgment of pupil performance with as much care as he devotes to the other aspects of his teaching task. In the case of the oral reading test, especially, the teacher should previously have assigned a value to each sound, or better, to each group

of sounds, and should endeavor to score each pupil as he reads aloud each scoring group or unit. It goes without saying that in classes in which progress in pronouncing and in understanding the spoken word is among the aims of the course, these scores should figure in the total estimate of pupil attainment; and it may be confidently predicted that teachers who endeavor, by trial and error, to develop a technique for testing oral and aural attainment will grow in skill and in ingenuity, and will, by the very fact of making these efforts, definitely stimulate their pupils to desirable growth in this direction.

2. Vocabulary

For many years the problem of acquiring a suitable vocabulary in the foreign language has been prominent in teachers' minds. It is probable that when Comenius published his topical vocabulary of Latin in 1631 under the title, *Janua linguarum reserata*, his choice of a title was symbolic of what most teachers of a foreign language have regarded as the core of their task. The advent of the Reform movement, with its condemnation of lists of words to be memorized — its partisans condemned with equal vigor the memorizing of declensions and conjugations — served in a way to cast discredit on the acquisition of vocabulary except in direct connection with some oral activity. But without discussion of this point, all must admit that progress in acquisition of vocabulary is an essential condition of progress in control of a language. Therefore, any picture of attainment in a language is incomplete without a wider sampling of vocabulary knowledge than we are accustomed to make.

Surely every teacher knows, on the basis of the textbooks used, what vocabulary items and idioms have been presented to his pupils, and, with the aid of recently accessible frequency

counts, which of these words and idioms rank high in usefulness. A vocabulary test and an idiom test of 75 or 100 items, or even more, prepared according to the multiple-choice technique exemplified in the American Council Alpha tests or the Cooperative Test Service tests, are not difficult to make and are very easy to score. The primary sources for data regarding the relative usefulness of words and idioms are, of course, the word and idiom counts developed by the Modern Foreign Language Study. These counts have now been supplemented by the Keniston Spanish list, the Purin German list (both issued by the University of Chicago Press), the French list of the Association of Modern Language Teachers of the Central West and South ("A Basic French Vocabulary," *Modern Language Journal*, January, 1934), and the lists in French, German, Italian, and Spanish printed in *A Tentative Syllabus in Modern Foreign Languages* (see above, note 3). Teachers may also wish to utilize from their particular courses items which play a sufficiently prominent rôle, although items that do not appear in these lists are rarely of sufficient importance to be included in a test.

Teachers who reject on principle such a test involving isolated words may prefer one consisting of sentences in the foreign language, each of which includes a word or idiom. The pupils must indicate the meaning of the word or idiom by choosing, from four or five possible items in English or in the foreign language, the one that comes nearest in meaning.

Example: Nous partirons *demain*.

today	aujourd'hui
at once	immédiatement
tomorrow	hier
hurriedly	le jour qui suivra celui-ci
yesterday	le jour qui a précédé celui-ci

EXAMINATIONS IN THE FOREIGN LANGUAGES

It is naturally more difficult to provide suitable foreign words or equivalents, while at the same time remaining within the assigned vocabulary limits.

Other techniques are also available for idiom and vocabulary tests. The familiar "recall" type, in which the foreign equivalent for an English word or expression is to be given, or the English equivalent for a foreign item, has the double disadvantage of not being objectively scorable in many cases, and of requiring more time for the same number of items. A useful variation of the recall technique is to give in the foreign language, or in English, a sentence from which a word or expression has been omitted. The pupil is to supply the proper word, either by inference alone or by selecting the nearest equivalent from a group of items. Examples are:

A. Pour voyager autour du monde on a d'argent.

1) Pupil by inference supplies *besoin*, or,

2) Pupil selects from beaucoup
 besoin
 trop
 lieu

B. To make a tour of the globe one *needs* money.:

manque
a besoin de
désire
veut

C. Les parents de la jeune fille sont mes

couleurs pères amis

It is possible for students to complete in not more than twenty minutes a hundred-item test of the kinds suggested, in which but little writing is to be done. And if the items are chosen on a difficulty scale, such as is represented by the arrangement of the items in Part II of the French and the Spanish Word Books, taking every fifth or eighth or tenth

item, but avoiding obvious cognates, the range of the pupil's vocabulary will be tested far more searchingly and fairly than in the conventional examination.

It may not be out of place to add that the kinds of test proposed, except the English-foreign language form of the "recall" type, measure above all the "recognition" or "passive" or "receptive" vocabulary, that is, the vocabulary for reading, the exact relationship of which to the vocabulary readily available in speech and in writing is unknown.¹⁷ But there is a relationship, quite certainly a positive one, and it would seem rather short-sighted not to take advantage of this relationship even though we may be unable to propose a variety of procedures for testing "active" or "productive" vocabulary directly.

3. The Cultural Element

We have canvassed some of the possibilities of testing with reference to at least three of the various aims which our conventional testing program does not take into account. A fourth is the so-called cultural aim. In the first place we hope that our modern language pupils will have a sufficiently increased interest to enlarge their stock of information by noting in the newspapers, magazines, and books they read, in the conversation of their associates, and in their other studies, informing or enlightening items bearing on the foreign country. They may also be led to read illustrative books in English. But all this must be reinforced by some purposeful and systematic utilization of what is contained in the course itself. Unfortunately, few attempts have been made to formu-

¹⁷ A tabulation of the vocabularies of seven "conversation manuals," prepared by Frenchmen for Americans, yielded a list of only 161 items common to the seven books. Of these, all but one had previously been included in the list entitled "A Basic French Vocabulary" (see above, p. 306) by the committee designated to draw up that list, and that one word is — *oie!*

late a canon for this aspect of the course at the different stages. What minimum of knowledge of France, Germany, or Spain — the country, the people — should we expect after two years of the language? What should a third year add to this? For thirty years, most beginners' books have contained views of Paris or of Berlin or of Madrid, pictures of cathedrals and short reading exercises that convey some information about geographic or other aspects of the country. The market is full of "daily life" books and elementary historical reading matter, but the subject matter of such books varies from one to the other, the material is used sporadically, and there is danger of its producing a "baby Baedeker," or a sort of course in history and geography to be studied as such.

The studies by Gilman and Kurz in French and by Van Horne in Spanish¹⁸ have shown conclusively that the literary texts read in French and in Spanish classes cannot be relied on to present any definite information or judgments in the "cultural" field, and there is no reason to believe that the situation is different in German. Consequently, teachers should first draw up a *modest* outline — and the adjective must be emphasized — of what they consider the minimum essentials for the course in hand, basing the outline on the textbooks in actual use, on the material in English, if any, which they expect their pupils to read, and on the information and the comments which they themselves intend to provide. The danger is, of course, that they will be over-zealous, and the results of this may be almost worse than the present casual attitude, or even than substantial neglect.

An outline of this sort will be useful to the teacher as a point of departure, but few of us can foresee precisely what we shall do in class. Consequently, the outline must be revised and

¹⁸ In *Studies in Modern Language Teaching*, Publications of the American and Canadian Committees on Modern Languages, pp. 225-63.

filled in as the days and weeks pass, until, at the end of the year, it has become the kind of bird's-eye view of the foreign country which is appropriate to the class in hand. In such a bird's-eye view for beginners, some very elementary historical and geographical data will appear; there will be glimpses of a few striking human figures and their deeds, of a few notable works of art, such as buildings, monuments, paintings; and the pupils will have had glimpses of the foreign people as they emerged into the light of recorded history and became what they are today. For the more advanced classes, this minimum will be added to in suitable proportions, until the teacher has developed a modest conspectus, or a little pamphlet, which records truthfully the specific information and judgments that have formed an integral part of the modern language course.

Needless to say, these topics would not be presented formally, in a series of lectures, but as the occasion arises; and the occasion *must* arise. The teacher should take notes so as to have an inventory of the points covered in a year or a semester. Otherwise he will not know on what topics he may ask questions when the time comes to set an examination.

Since most of what the pupils acquire in this field will be informational in character, the examination questions may readily take the objective form, calling for a check or a choice between items. Furthermore, as most information must be conveyed orally through the medium of English, the questions should preferably be in English, although the inevitably greater ease with which students read a foreign language as compared with their ability to understand it and to speak it will admit of some freedom in this respect, provided they are allowed to write their answers in English, when writing is required by the technique adopted.

4. Reading

The problem of how to test the reading ability of pupils at any stage is, in many respects, one of the most difficult, and this fact explains in part the deficiencies in the conventional examination which have been noted above. Experience has shown that correlations between reading ability and vocabulary scores are higher on the whole than between reading ability and any other language ability as measured by tests that have been administered on a large scale.¹⁹ But the tendencies, as evidenced by the data now available, are not pronounced enough to absolve teachers from the obligation of measuring reading ability more accurately than is commonly done. On the other hand, Henmon's study of the correlations between teachers' marks and attainment in the various abilities, as measured by the American Council Alpha tests,²⁰ shows clearly that reading ability usually plays a definitely smaller rôle in determining class standing than knowledge of grammar, thus indicating that the grades given by teachers in the United States have a relatively small value as indices of reading ability. Data from Canada cited by Professor Henmon indicate that Canadian teachers take more account of reading ability in grading their students, although in the same volume (pp. 262-66) other data demonstrate that Canadian pupils read French relatively less well than do high-school students south of the border. A satisfactory measure of reading ability should reveal, among other things, both the rate of reading and the degree of comprehension. In other words,

¹⁹ V. A. C. Henmon, *Achievement Tests in Modern Languages*. Publications of the American and Canadian Committees on Modern Languages, vol. v. New York: The Macmillan Co., 1929, pp. 92-97.

Frederic D. Cheydleur, "Attainment by the Reading Method." In *Experiments and Studies in Modern Language Teaching*. Chicago: University of Chicago Press, 1934, pp. 100-44.

²⁰ Henmon, *op. cit.*, pp. 97-103.

it should, if the word "reading" is thought of in its customary meaning, reveal how closely an individual's reading performance in the foreign language approaches the way he reads in the mother tongue. To be sure, one may read a passage in the vernacular in various ways according to the purpose in hand.²¹ It is almost certain that, in the great majority of cases, pupil reading in a foreign language is of the "study" type and only rarely approaches the desired norm represented by behavior in reading similar material in the vernacular. But it must also be observed that the pupils' reading experience in the foreign language tends to promote this type of result.

It is evident that the usual reading passage in an examination, whether pupils are required to translate it or answer questions based on its content, throws no light on rate of reading. The efficacy of "speed drills" or practice exercises for increasing the reading rate has been pointed out often in recent years by authorities on the development of improved reading ability in the vernacular,²² and exercises of this type were used with profit by Michael West in teaching Indian children to read English.²³ These consist of a suitable passage or a group of connected paragraphs on which questions are based. Professor Gates, in the volume already cited, gives an interesting discussion of ways in which reading in the vernacular may be improved (pp. 205-17), and offers specific examples of devices by which comprehension may be tested.

Teachers will find it useful to select for each fortnightly or monthly classroom quiz five to ten short passages, which they

²¹ C. H. Judd and G. T. Buswell, *Silent Reading: A Study of the Various Types*. University of Chicago Press, 1922.

A. I. Gates, *The Improvement of Reading*. New York: The Macmillan Co., 1927, pp. 180-81.

²² See, for example, Gates, *op. cit.*, pp. 206, 227-30.

²³ Michael West, *Bilingualism*. Calcutta: Government of India, Central Publication Branch, 1926, pp. 202-15.

think differ in difficulty, and to set as many questions on each passage as seems appropriate. The performance of pupils on each reading test so made up will provide a basis for deciding which of these passages and questions are best suited to the purpose. A record should be kept of the scores on each passage making up the quiz. As a result, the teacher will be able, at the end of the semester or year, to select from the passages and the questions those which have best served the purpose in hand, and thus to construct out of them a final reading examination, the component parts of which have been subjected to a trial administration.

Such a procedure has two distinct advantages. It enables the teacher to eliminate questions which, in the light of experience, have turned out to be ambiguous or which have failed to test understanding of the passage. Most standardized reading tests of the paragraph-question type have been constructed as the result of following this general procedure, only, of course, on a larger scale. Furthermore, administrations of this kind will enable teachers to judge better the relationship of the time allotment to the number of running words in the whole test and to the number of questions. It was found in the case of the American Council Alpha French reading test, which was administered to pupils at all semester levels (1 to 8), that an allowance of about 32 minutes was adequate for a test of seven passages, containing a total of approximately 1000 running words, on which 28 questions were based. Not all pupils were able to complete the test in the time allowed, partly because the passages increased in difficulty and partly because of a slow reading rate. But since the element of speed is an essential factor in reading, it is entirely appropriate that a reading examination should display some power of discrimination in this respect. At the same time, it is clear that an examination which, as in the case mentioned, demands a reading

rate of fewer than 50 words per minute, can hardly be charged with undue "speeding-up."

If a teacher has become sufficiently interested in developing reading ability to devote a specific portion of the class hour to practice in reading and to testing progress — which, unfortunately, few teachers do — the stage is set for a test at the time of the semester examination, which, by utilizing the same sort of technique, will more nearly reflect the ability of pupils to read and understand in the usual sense of the word, in contrast to an exercise in matching equivalents or to a display of grammatical knowledge, such as characterizes a translation exercise. Indeed, while a discussion of the question is not pertinent here, we must take cognizance in passing of a growing conviction among investigators in the field, not only that one cannot measure reading ability through translation, but that regular classroom practice in translation, in order to prepare pupils for a translation test, has a positively deleterious influence on progress in reading, if we are to interpret the word "reading" in its proper sense.

The vocabulary, idiom, and grammar sections of the testing program should throw ample light on attainment in these particulars. If we assume that these are adequately cared for, the only important rôle left to the translation passage is to test (a) whether the pupil is familiar with the similarities and differences between sentence patterns in English and in the foreign language, and (b) whether he is competent to express in his mother tongue what the author has adequately said in his. To anyone familiar with school compositions in English and with school versions of a foreign original, it is clear that the unsatisfactory quality of a translation may easily be due to other causes than inability to read the foreign language. And when, in addition, we take into account the fact that it is virtually impossible to score the translations made by a class

in such a way as to reflect accurately the relative performance of each individual, we are forced to conclude that this method of measuring ability to read a foreign language has but little in its favor.

In order to test the reading ability of grade school pupils in English, Professor Gates devised tests ²⁴ to measure:

1. Comprehension of the general significance of a paragraph.
2. Ability to predict the outcome of events as stated in a paragraph.
3. Ability to understand precise directions.
4. Ability to note details.

Now it is unlikely that the modern language teacher needs specific tests for each of these special purposes. At the same time, he would find it useful, when constructing a reading test, to keep some such analysis in mind so that the questions he puts may measure the pupils' progress.

Since limitations of space forbid a complete illustration, let us take as an example one paragraph from the American Council Alpha French test and suggest questions in conformity with Gates's proposals. The questions may be put in English or in the foreign language, but in the latter case, the vocabulary and syntax must lie within the range of the pupils' experience.

Un pauvre aveugle n'avait que son chien. Un jour on lui vola ce compagnon fidèle et il resta seul. Il l'appelle: il se sent aveugle deux fois. Un passant s'approche de lui et demande: "Mon ami, qu'y a-t-il pour votre service?" L'aveugle lui raconte son malheur. Le monsieur se met en colère. Il dit: "Je suis juge. Si je peux trouver le voleur, je jure de le faire punir." L'aveugle dit: "Monsieur, la vengeance ne me donnera rien. Je ne veux punir personne. Je vous prie seulement de me faire rendre mon chien."

²⁴ *Ibid.*, pp. 184-91.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

1. Which of these words tells best how the passerby felt?
blind deserted angry joyful
2. Which of these sentences tells what you think happened next?
The dog came back to his master.
The man had the guilty person put in prison.
The gentleman had the police look for the dog.
The man insisted that he must have revenge.
3. Lesquelles des phrases suivantes disent ce qui est vrai?
Le passant est aveugle et sans son chien il se sent aveugle deux fois.
L'aveugle s'approche du passant.
L'aveugle demande au juge de faire punir son chien.
Le monsieur veut faire punir celui qui a volé le chien.
L'aveugle dit qu'il ne désire pas se venger.
L'aveugle veut

{	punir le voleur.
	retrouver son chien.
	aller chercher le juge.
4. Qui se met en colère? Le chien L'aveugle Le passant
Qui parle le premier? Le voleur Le passant
L'aveugle

In actual practice the teacher would choose paragraphs of differing types, each of which would lend itself to one of the kinds of testing that he had in mind. He would take care that the actual subject matter be new, but that the paragraph be in harmony with the reading program of the class and contain no words that were new to the pupils; or, if new ones did occur, that they should not be key words to the meaning of the paragraph or should be so used in the text as to be understood by inference. In this respect the teacher who is testing a particular class or classes has an advantage over those who construct reading tests for wide administration in many schools in which course content may differ greatly. It may be added that there is distinct value in a reading test so constructed as

to ascertain whether pupils can understand new words by inference, whether these are or are not cognates.

Before the semester- or the year-end examination has been reached, teachers should have ascertained roughly the reading rates of their pupils in terms of running words per minute. One would expect the rate to be higher in classes reading texts which have a graded "vocabulary burden," that is, a limited number of new words to each 100 running words, than in classes using ungraded texts. If, for example, students have been trained to read in graded texts involving only, let us say, the first 2000 items in *French Word Book*, Part II, and then are tested as to reading power by two such passages as are found in the Cp. 2 French Examination for June, 1933, they will find in the first passage of about 285 running words one word (*au-dedans*) that does not occur in *French Word Book*; five (*fabuleux, ampleur, natal, clientèle, écolier*) which belong in the fifth thousand; one (*héritage*) in the fourth thousand; four (*plaisanter, sud, botte, douzaine*) which fall in the third thousand. It is likely that *douzaine, fabuleux, clientèle, and héritage* would be understood by many because of their resemblance to English in form and meaning — the last two have a low merit index in Thorndike — but this is not likely to happen in the case of *botte, au-dedans, sud, ampleur, natal, plaisanter, écolier*. In the second passage, of about 210 running words, to be translated, three words (*geler, boue, os*) belong to the fourth thousand of *French Word Book*; one (*charrette*) to the fifth thousand; and one (*espionnage*) is not found in that list. It is, in a certain sense, a common word, or has been so especially since 1914, but it did not appear in as many as five of the 88 texts which provided the material for *French Word Book*. Consequently, the rate and accuracy of reading as measured by questions on two such passages would vary according to whether the questions asked involve knowledge of

the precise meaning of these words. In this particular case, it is highly probable that the five words in the second passage which we have assumed to be unfamiliar will lower the pupils' scores more than the eleven unfamiliar ones of passage one, because, of course, of the particular demands made by a translation test.

5. Grammar

The devotion of members of the profession to grammar has produced at least one desirable result. Our techniques for testing grammar knowledge are varied and ingenious. The exercises of almost every recent textbook reflect this result.²⁵ Certain of the types of grammar question proposed by the Committee of Twelve are rarely met with nowadays. Examples from the Committee's "elementary" examinations are:

Write a synopsis of the conjugation (1st person singular of each tense) of *se réjouir* and *savoir*.

Write the forms of the demonstrative pronouns.

Decline throughout the German phrases meaning *the new house*, *my dear friend*.

Give the third person singular of each tense in the indicative mood of *bittend*, *blieb*, *schlug auf*.

On the other hand, certain kinds of traditional tests of grammar knowledge are still in favor; sentences for translation into the foreign language which bristle with difficulties for the tyro, as, for example, those involving such puzzles in French as "Ce à quoi je pense" or "Ce dont je parle" or "Le monsieur avec le fils de qui j'ai voyagé," and connected passages introducing similar puzzles which are to be translated into the foreign language.

²⁵ For useful specimens of techniques in constructing examinations in grammar, vocabulary, and "comprehension," see G. M. Ruch and G. A. Rice, *Specimen Objective Examinations*. Chicago: Scott, Foresman and Co., 1930, pp. 224-48.

Now, even if due precautions are taken about relating the demands on vocabulary and idiom to the pupils' experience, one can still object seriously, on one score at least, to such an examination device as the last. It is virtually impossible to grade the test uniformly unless the teacher takes elaborate precautions, as is done by the readers of the College Entrance Board.

Some examples copied from pupils' papers in a recent second-year examination in German illustrate the point:

1. I cannot wear these shoes; they (demonst.) are not large enough.
Ich kann nicht diesen Schue tragen; der sind nicht ganz genug. (—1)
2. The street on which he lives is the most beautiful in the city.
Die Strasse worin er liebt, ist das schönste in den Stadt. (—1)
3. What is the name of the people with whom he lived?
Was ist die Name von der Laude, womit er liebt? (—1)
4. Some things (manches) she says are true.
Manches sie sagt, sind rein. (—1)
5. What pieces of music does she prefer?
Was fur stucke music wunscht sie Lieber? (— $\frac{1}{3}$)
6. Whoever studies industriously will not fail.
Wer einzeln studiert, wird nicht — (— $\frac{2}{3}$)
7. You (Du) who told me that know still more.
Du, der du das mir erzählen hast, weisst jetzt mehr. (—1)

The numbers in parentheses preceded by the minus sign indicate the deduction made by the teacher on each sentence. The section, valued at 25, was composed of twelve such sentences, and the examiner evidently attempted to give each error a suitable weighting. One has only to compare the deductions on numbers 4 and 6, and to consider that each sentence offers different possibilities of error for each pupil, to realize how vain such efforts are. A variorum edition of the

pupils' translations of sentence number 3, with the teacher's deductions, illustrates the point:

Was ist die Name der Leute, bei denen er gelebt ist? ($-\frac{1}{3}$)
 Was ist die Name des Leutes mitwem er lebte? (-1)
 Wie heissen der Leute womit er wohnt? ($-\frac{1}{3}$)
 Was ist die Name dieses Man bei — er wohnte? (-1)
 Was ist der Name, bei den Einwohnern, womit er legte?
 ($-1\frac{1}{3}$)

As every teacher can add dozens of examples each time that he reads a set of test papers which contains a section of this sort, it is unnecessary to insist further on this point. Practice in translating such sentences into the foreign language (although one might object that a sentence like number 7 is too foreign to the pupils' way of thinking, speaking, and writing) has its values in the second year and thereafter, but as a device for examining, little can be said in its favor.

Similar difficulties in scoring arise, of course, when a longer English passage is given to translate, or when "free" composition is called for.

The implication from the preceding paragraphs is that a grammar test should conform rather closely to the type exemplified in the standardized tests mentioned above. Each phenomenon to be tested should be isolated in such a way as to simplify the scoring. The devices for achieving this are numerous and familiar, and specimens are to be found in most school grammars of recent date. Some of them are:

1. Require the pupil to complete a sentence by selecting the correct word from three or more which are proposed:

Le frère et la soeur sont (parti, partis, parties) hier soir.

2. Require the pupil to select from several proposed forms the form which renders correctly a given English sentence or phrase. Example:

EXAMINATIONS IN THE FOREIGN LANGUAGES

Directions: Draw a circle around the number of the correct translation of the English sentence: He has been here for a week.

1. Il a été ici depuis une semaine.
2. Il est ici depuis sept jours.
3. Il est ici depuis huit jours.
4. Il a été ici pendant une semaine.

3. Require the pupil to supply what is missing or eliminate what is superfluous in order to make a sentence formulate accurately a statement of grammatical form or usage:

1. The possessive adjective does not agree in gender and number with the thing possessed.

2. Reflexive verbs are conjugated with the auxiliary

4. Formulate several grammatical "rules," number or letter them, and then ask the students to indicate by the proper number or letter the one of these "rules" that is illustrated by each of a list of sentences or phrases. Example:

1. The future tense form must be used in subordinate clauses when futurity is implied, especially after *quand*, *lorsque*, *aussitôt que*, *dès que*.

2. The possessive adjective has the gender and number of the thing possessed.

3. A personal pronoun agrees in person and number with the noun for which it stands.

Mon père a perdu sa plume.

A qui cette montre? Elle n'est pas à moi.

Venez me voir dès que vous serez guéri.

5. Require the pupil to shift the words of a given sentence, e.g. the tense or mood of the verbs, the number of the nouns and adjectives, the degree of the adjectives, etc. Example:

1. Put the verbs in the future tense:

Llegamos a una casa de la calle de L. Don José no está en su domicilio.

2. Change to the plural where appropriate:

Soy malagueño. Cuando chiquitín quería ser marino.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

3. Replace the infinitive by the proper form of the subjunctive:

Quiero que todos junto (hacer) este viaje.

6. Require the pupil to substitute object pronouns for nouns, to test knowledge of forms and of word order.
Example:

Substitute *y* or *en* or personal pronouns for the italicized words: Il demande *la plume à sa sœur*.

7. Give a numbered list of tense names, in English or in the foreign language, and require that the items in an accompanying list of verb forms be identified by the proper number. Example:

1. Present subjunctive	Nous donnâmes
2. Pluperfect indicative	Ils eussent
3. Past absolute indicative	Qu'il boive
4. Past subjunctive	Nous nous étions endormis.

8. Ask for a translation of an English word or expression so as to complete a sentence in the foreign language:

Alonso lee (less than) Vd. se imagina.

9. Require the pupil to select what is necessary to complete a statement in regard to grammatical usage. Example:

Directions: Check in the proper column.

	ser	estar		
1.	—	—		
2.	—	—	is used to express {	
3.	—	—		an accidental state
4.	—	—		nationality
5.	—	—	is used to {	ownership
6.	—	—		location
				form the passive voice
				express a permanent condition

Other devices could be enumerated, but these are sufficient to illustrate the point under consideration. The techniques suggested under items 1, 2, 7, and 9 in the preceding list may

be regarded as testing primarily a "recognition" or "receptive" knowledge of grammar, and are therefore especially suitable to the testing program of the first two years. It is an established fact that examinations which utilize techniques of this sort yield somewhat higher scores than tests in which the pupil must manipulate the language. In other words, such techniques make it possible for the pupils' knowledge to "register" over a wider range of phenomena. Furthermore, such techniques are undoubtedly more in harmony with the kind and the degree of grammatical knowledge possessed by pupils at this stage of language study.

The objection may be made that such situations are artificial, and that requiring a translation of English sentences, so arranged as to put the student entirely on his own resources in using the proper form of expression, is a much more searching test. This objection has weight, but it has been largely forestalled in what has been said above. In the first place, it is well-nigh impossible to score a test of this kind with accuracy, and in the second place, sentences of this sort usually become grammatical puzzles to be solved, and so do not represent anything that the student would normally say or write in the language if allowed to express himself in his own way.

6. Composition

Reference has already been made to the difficulties encountered in scoring "free" compositions and translations of English passages into the foreign language. Differences of 10 to 40 points between scores on the same passage given by different instructors are not uncommon. It is clear also that the same instructor scoring different passages is in danger of varying almost as widely in grading the individual members of his class.

The use of composition scales developed as a part of the test-

ing program of the Modern Foreign Language Study serves to render scoring on "free" compositions more nearly uniform.²⁶ Professor Ford ascertained that, if students in advanced classes are required to select two subjects for ten-minute compositions out of four proposed, and if these are scored with the aid of a composition scale, the resulting scores are much more reliable than if teachers read the compositions, noting all the errors, and then attempt to assign a value to each error or to estimate a suitable deduction of a given number of points for the total number of errors. He found, also, that extremely reliable individual measurements may be arrived at by requiring such compositions on four out of eight possible subjects. The two experimental groups in this case were made up of students who had taken French for five and for six years respectively. The two groups of subjects were:

A

1. A friend has asked you to go to the theater. Explain why you cannot go.
2. The French book that you enjoyed most during your first year.
3. The first week at the university.
4. An exciting adventure.

B

1. The advantage of a knowledge of French.
2. A scene or happening in domestic life.
3. A scene or happening on a journey.
4. The city of Toronto.

²⁶ *Modern Language Instruction in Canada*, 1. Publications of the American and Canadian Committees on Modern Languages, vol. vi. University of Toronto Press, 1928, pp. 506-31.

V. A. C. Henmon, *op. cit.*, pp. 34-62.

Frederick S. Breed, *Studies in Modern Language Teaching*, pp. 187-98.

H. E. Ford, *ibid.*, pp. 201-10.

As readers who are familiar with the use of a composition scale know, the scorer, instead of marking each error in the composition, attempts to judge its *general merit* by comparing it as a whole with the different specimen compositions of the scale, which have been ranked in order of merit as the result of an elaborate procedure based on the judgment of experts; and by giving it the rank order of the scale passage to which it is nearest in general quality, under which term should be included length, extent and appropriateness of vocabulary, spelling, conformity to grammatical usage, and the like.²⁷

It appears to have been the custom of such agencies as the Secondary Education Board to assign three or four composition subjects in French from which the candidate is to choose one on which to write 50 words or more, according to the stage of advancement. In the 1933 booklet,²⁸ one subject is indicated for French I, and three each for French II and III, from which one is to be selected. These are:

French II: Une visite chez un ami.

Mon père.

Une leçon de français.

French III: Un portrait de vous-même: apparence, goût, caractère.

La journée d'un maître d'école.

Une promenade avec mon père ou quelqu'un avec qui j'aime à me promener.

It is evident that to adopt the plan which Professor Ford's investigations found to be most satisfactory, the time allowance for the composition section must be a minimum of 20 or

²⁷ The composition scales in French, German, and Spanish which were developed by the Modern Foreign Language Study form a part of the American Council Alpha tests, which are published by the World Book Co., Yonkers-on-Hudson, New York.

²⁸ *Definition of the Requirements for 1934, with the Examinations for 1933.* Office of the Board of Secondary Education, Milton, Massachusetts.

of 40 minutes, according to whether the pupils are required to write on two out of four subjects, or on four out of eight.

While the illustrations given here are based on tests of writing ability in French, it goes without saying that the same principles are applicable to German and to Spanish.

In the opinion of the Committee responsible for this chapter, tests of ability to compose consecutively in the foreign language are more appropriate at the third- and fourth-year levels than earlier, and, if an adequate sampling is desired, may well occupy a class hour prior to the period allotted for semester- or year-examinations.

7. General Linguistic Knowledge

Nothing has been said so far about testing the students' increased knowledge of language in general, and especially of the English language. The content of such a test will, naturally, depend on the treatment of this topic in class. In French and in Spanish, the use of prefixes and suffixes promptly suggests itself: *-ment* and *-mente* to form adverbs; *im-* with the same value as in English; *-eux* and *-oso* in adjectives with the value of English *-ous*; *-esque* and *-esco* equal to English *-esque*; *re-* with the value of English *re-*.

The numerous cognates in the two Romance languages and in English loom very large. Of these, the less obvious ones are frequently the most instructive. A few such French-English cognates are:

cheval and cavalier, vendre and vendor, meilleur and meliorate, main and manual, parler and parley, vanter and vaunt, matin and matins, enfance and infancy, deviner and divine (verb), rêve and revery, maison and mansion, pauvre and pauper, colère and choler, atteindre and attain, corps and corpse, courant and current, divers and diverse, doigt and digit, école and school, environ and environs, étage and stage, façon and fashion (way), faute and fault, fou (folle) and folly, gros

and gross, larme and lachrymal, manteau and mantle, mode and modish, morceau and morsel, nouveau (nouvel) and novel, parole and parole, propre and proper, régler and regulate, sauter and somersault, tailler and tailor, terre and terrestrial, vide and void, caisse and cash, prévenir and prevent.

Since these French words are found in the "Elementary" word list of the New York State syllabus referred to above, and all but one (*vanter*) in "A Basic French Vocabulary" (*Modern Language Journal*, January, 1934), it is evident that the teacher of French does not have to ransack the dictionary in order to find material for the purpose in hand.

Examples of the relationship of English to German, both in word formation and in cognates, are so numerous as to be almost embarrassing.²⁹ Indeed, the danger here is no less than the danger already pointed out as confronting the teacher and the examiner with respect to the "cultural" element. The content of the foreign language course must not be converted largely into an assemblage of facts drawn from history, geography, literature, and the like, nor should it become to a considerable degree an elementary course in comparative linguistics. But if development of a knowledge of linguistic relationships is a valid and desirable objective, as virtually all pronouncements on the subject agree, specific material must be utilized to promote progress toward it, and it is only reasonable that the progress made should be measured as a part of the general accomplishment.

One obvious procedure in doing this would be either to give a list of suitable foreign words chosen from the class list and to

²⁹ German grammars often provide material. See, for example, the vocabulary of E. Prokosch, *Deutsche Sprachlehre*, Holt, 1930; Peter Hagboldt and F. W. Kaufman, *Lesebuch für Anfänger*. Chicago: University of Chicago Press, 1930.

See also Peter Hagboldt, *Building the German Vocabulary*. Chicago: University of Chicago Press, 1930, pp. 71.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

ask for the English cognate, or to do the reverse. Or one might follow the example of Woody and Hootkins in the study on growth in vocabulary by students of French, which has been referred to above, and give a series of English sentences each containing a cognate, the meaning of which is to be indicated by the pupil. Two examples:

<i>Manually</i> , he is quite capable.	{	usually as an athlete mentally naturally with his hands
Everyone agreed upon a <i>parley</i> .	{	auction conference execution game trial

Or one might ask pupils to identify all the cognates to be found in one of the passages assigned in the reading test. Or one might call for a differentiation between pairs of cognate forms and their current meanings. Illustrations of this are:

mansion — maison; harvest — Herbst; alumnus — alumno

Tests of knowledge of the uses of prefixes and suffixes in the different languages are readily constructed.

Finally on this topic, a word of warning to those teachers who overrate the familiarity of pupils with their mother tongue and who lean heavily on cognates as a key to the foreign language, only to discover, to their dismay, that the vernacular word is often almost as strange to their pupils as the foreign word. The investigations of Limper and of Dale³⁰ make it clear that it is easy to assume too much in respect to the know-

³⁰ Louis H. Limper, "Student Knowledge of some French-English Cognates," *French Review*, VI (November, 1932), pp. 37-49; Edgar Dale, *Familiarity of 8000 Common Words to Pupils in the Fourth, Sixth, and Eighth Grades*. Bureau of Educational Research, Ohio State University. Unpublished.

ledge of slightly uncommon English words possessed by very many pupils. Teachers can protect themselves from this danger, at least in part, by making sure that the vernacular words have a fairly high merit index in a standard English word list. As an illustration, we may note that the English cognates of the two French words *clientèle* and *héritage* which were referred to above (page 317) rank low in Thorndike's *Teachers' Word Book*.

If we now attempt to suggest the component parts of an actual semester- or year-end examination, we must assume that few schools allow more than two hours for this purpose. A two-hour period would allow some such program as the following:

1. A 100-item vocabulary and idiom test by a multiple-choice or some other quick-answer technique, 15 to 20 minutes.
2. A 50- to 75-item grammar test, similarly devised for quick answers, 30 to 35 minutes.
3. A test of information and judgments about the foreign country and people, constructed for short answers, 10 minutes.
4. A reading test of eight to ten passages, with four to eight questions or problems under each passage, 50 to 60 minutes.

Such an exhibit reinforces what has been indicated above; namely, that provision must be made prior to the examination period proper for measuring pupil progress in oral and in aural abilities, in speed of reading, and in composition. In schools in which even less than two hours is allotted to regular examinations, such an examination as is proposed above may be administered in suitable sections on two or even on three successive days. The important consideration is, of course, that pupils should be adequately tested as near as possible to the end of a scholastic time unit rather than that the test should come on a given day.

The distribution of the examination time as suggested above is probably more appropriate for first- and second-year classes than for more advanced groups. In the latter case, when attainment of the reading attitude has once been ascertained, teachers may well prefer to give a larger relative share of the time to items 2 and 3, and to tests of ability to write in the foreign language. But because of the importance of measuring progress toward reading ability, and in view of the virtual futility for that purpose of the conventional examination, the time allotted to that test unit in the preceding proposals is most certainly not out of proportion.

It has already been suggested that the practice exercises for improving the reading rate, if timed with fair accuracy and checked by content questions, provide the best means both of developing and of measuring speed. Not all the members of a given class will need to be included in all the practice exercises, because some will have become efficient before the others; but all should be included in the exercises specifically designed to test speed as the end of a scholastic period draws near.

Perhaps the chief emphasis in what precedes has been placed on examinations to be administered at the end of a semester or a year, and perhaps these are the most difficult to construct in accordance with the principles laid down in this volume. Ideally, all such examinations should be constructed in harmony with the course of study by, or with the aid and advice of, the research bureau of a school system, or by a committee of teachers who have carefully studied testing procedures. Individuals who see fruitful possibilities in these pages are urged to try out repeatedly in brief daily, weekly, and fortnightly quizzes the principles and techniques suggested. The practice thus gained in choosing materials and techniques will stand them in good stead when devising longer and more important examinations.

CONCLUSION

In closing this enumeration of the various ingredients of a testing program designed to present a more accurate picture of the progress of students toward their objectives than is provided by the conventional examination, it is well to emphasize what has already been suggested. The chief effort at the stages of modern language study with which we are chiefly concerned must be to acquire knowledge of and skills in the language itself. Consequently, the portions of the examination devoted to testing progress in the "cultural" aspect and in comparative linguistics, including growth in knowledge of English, should occupy a subordinate rank as compared with the other portions, in harmony with the relative amount of time devoted to the various topics in the classroom.

An uncritical reader might hastily and erroneously infer from what precedes that testing in modern languages is more defective than in other subjects, and that the poor classification of students referred to above is peculiar to this field. The data given by Professor Gates³¹ with respect to (a) the percentages of failures in grades 3 to 8 caused by deficiencies in reading English, and (b) the wide variations in reading ability in English in six New York schools, provide evidence of equally great disparity in the attainment of such a fundamental aim as the ability to read with comprehension at a fair rate of speed in the vernacular. It has been shown also that many college students are handicapped by poor attainment in reading ability.

Modern language teachers have not, on the whole, been more unsuccessful in measuring the results of their teaching

³¹ *Op. cit.*, pp. 4-6, 199-200. Cf. for mal-classification in physics, Ben D. Wood, "The New York Experiments with Modern Language Tests," *Publications of the American and Canadian Committees*, vol. I. New York: The Macmillan Co., 1927, p. 185.

than their colleagues in other subjects. But when confronted by the state of affairs prevailing in their own field, they cannot afford to console themselves by contemplating equally inadequate efforts to measure attainment in other departments. It should be evident that more thought ought to be given to an analysis of the testing program and its relationship to the aims and the content of the course, that more effort should be expended on the construction of the tests themselves, and, finally, that good testing is an integral part of good teaching.

QUESTIONS FOR DISCUSSION

1. What is the relationship between the aims of a modern language course and the examinations set at the close of a semester or a school year?
2. What relationship exists between the conventional testing program and the state of affairs set forth in the works referred to on p. 293, note 5?
3. Reword the five statements on pp. 294-95 so as to make them affirmative instead of negative (e.g. "all examinations . . . test progress in . . .")
4. Assume that the statements in this new form should become true. What changes would *ipso facto* result in the modern language field?
5. Secure copies of recent examinations set by the C.E.E. Board and the New York Regents. Analyze them, and compare the results with the analyses given on pp. 299-300.
6. To what extent are these examinations free from the deficiencies enumerated on pp. 294-95?
7. What difficulties does one face in making tests of ability: (a) to understand the foreign language aurally; (b) to pronounce the language satisfactorily; (c) to use the language in speech?
8. Why is it possible for a teacher to judge more effectively of the aural and oral progress of a class than for someone who must rely entirely on the results of a test?

EXAMINATIONS IN THE FOREIGN LANGUAGES

9. Endeavor to draw up lists of the phonetic phenomena in your principal language on which it would be appropriate to test pupils at the end of the first, the second, the third years of study.
10. Using the techniques suggested in paragraphs a and b (p. 303), make two specimen ten-item tests of aural recognition.
11. Examine as many vocabulary tests as you can assemble (standardized tests, classroom tests) and report (a) on the character of the words chosen, (b) on the techniques used by the authors of the tests.
12. Using a first-year or a second-year textbook, make five ten-item vocabulary tests, exemplifying the five techniques described on pp. 306-07.
13. On the basis of the content of a given course of study (grammar text and readers) for a given semester or year, make a list of the "cultural" items which appear.
14. Consult the article by J. B. Tharp, *French Review*, VIII, 4 (March, 1935), 283-87, and chapter VIII of Florence M. Baker, *The Teaching of French* (Houghton Mifflin, 1931) as a point of departure in making an examination with the materials so gathered (no. 13 above).
15. What disadvantages are inherent in measuring ability to read on the basis of an English translation of one or two passages in the foreign language?
16. Study the discussion by Professor Gates of testing reading ability in the mother tongue (see the reference on p. 315) and point out the ways in which his procedures apply to a foreign language.
17. Select four appropriate passages from the reading texts of a given course, and utilize for each passage one of the testing techniques illustrated on pp. 315-16.
18. With these same passages, make a reading examination using the paragraph-question technique as in the American Council Alpha tests.
19. Construct a variant of this test by the use of the "best answer" technique. See Baker, *op. cit.*, p. 221.
20. If you have access to classes in the foreign language for ex-

perimental purposes, make an attempt to ascertain the pupils' reading rate. Select an unfamiliar passage of about 250 running words in length, ask each pupil to read this silently and to raise his hand and close his book when he has read the passage so that the time elapsed may be noted. At the end of five minutes direct that all books be closed and ask the pupils to write the answers to questions previously written on the board but hidden by a map or a newspaper. A passage of this length would be suitable for a five-minute test during the second semester. A more advanced group should have a longer passage or a shorter time allowance, say 3 or 4 minutes. The passage, while unfamiliar, must be written in a familiar vocabulary. Administer such a test fortnightly during a semester, utilizing for the last test, the one first administered. Plot the curve of progress in terms of running words per minute, and of the percentage of correct answers to the questions.

21. On pp. 320-22 are illustrated nine techniques for constructing objectively scorable grammar examinations. Make a group of five-item grammar examinations which exemplify these various techniques.
22. What are the weaknesses of *testing* knowledge of grammar by having pupils translate English sentences or passages into the foreign language? What are the virtues of such exercises for *teaching* purposes?
23. Secure a half-dozen or more "free" compositions written by pupils. Have copies made, and, in collaboration with several members of your class, have each collaborator grade these papers in the usual way. Then, after an interval of several days, have each collaborator, or a different group of collaborators, score them with the aid of the appropriate composition scale belonging to the American Council Alpha tests, and compare the scores so arrived at. This should be done in the case of several sets of composition. No marks should be made on the papers, and collaborators should work independently after making sure that all understand the problem and will proceed in the same way. It would be well to read the articles by Ford and Breed (see p. 324, note 26). Which grading procedure yields the more nearly uniform results?

EXAMINATIONS IN THE FOREIGN LANGUAGES

24. Assemble from a given text a list of the less obvious cognates as illustrated on p. 326, and construct a group of tests of five items each in which you exemplify the techniques presented on p. 328, and any other techniques which you devise yourself.
25. Make a list from a given textbook of the prefixes and suffixes so commonly used in the vocabulary of that text that pupils should learn their values and be tested upon them.
26. Of what units would a complete battery of tests of progress in a modern language be composed?
27. Study the Spanish test given in Ruch and Rice, *Specimen Objective Examinations*, pp. 229-34. Utilizing the same techniques, make, on a reduced scale, a similar examination, covering three or four lessons of a first- or a second-year book. Not all portions of the examination can be adapted for French and for German.
28. In constructing a reading test of the "paragraph-question" or "best answer" type, there is always a possibility that ingenious pupils may be able to give the right answers without having read the passages. They can do this by inference from the general character of the questions or statements, or from the phraseology of the erroneous responses, or because the questions or statements are so detailed and follow so logically that one item gives a clue to a preceding item, etc. For example, in the Spanish Reading test in Ruch and Rice, pp. 238-41, there are 40 items based on a brief passage, pp. 236-37, which entails a very close analysis; and without looking at the Spanish text, one could hardly miss numbers 4, 5, 6, 8, 11, 12, 13. Try out all items in the test in the same way, mark your score, and compare it with the key on p. 247.
29. Similarly, one can answer correctly the first question on p. 244 of Ruch and Rice with the aid of question 2, and with the aid of question 6 one can answer correctly questions 4 and 5, and probably question 7. Examine all the items of this test in the same way, and check your score by the key (p. 248).
30. Select a passage of 100 to 150 running words and construct a test according to the technique exemplified on pp. 244-47, of Ruch and Rice, but much briefer. Then have a classmate sub-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

- ject your examination to the same kind of scrutiny as above. Compare the reading tests constructed according to nos. 18 and 19 above.
31. If the examination program outlined on pp. 329-30 is in your judgment and experience impracticable for the school situations with which you are familiar, make up a program that will be more appropriate. Defend your position.
 32. Read in *Modern Language Journal*, *French Review*, *German Quarterly*, *Hispania*, *Bulletin of High Points* (New York City) the articles that have appeared since 1932 on making examinations and summarize what you learn from them.
 33. Secure from several schools in your vicinity specimens of their semester examinations. Report on the techniques employed and on the aims which the makers of these examinations appear to have had in mind.

CHAPTER VII

EXAMINATIONS IN MATHEMATICS

THIS chapter must not be considered an isolated and self-contained unit. It is intimately related to its wider context, and takes much of its meaning from that context. More particularly, it has to do largely with specific applications, in the province of mathematics, of those broader rules and principles laid down in Chapters II and III, and the reader can hope to see the present discussion in its proper perspective only if he has first made himself familiar with that more general treatment of test construction. Throughout this chapter a background of such familiarity will be assumed.

TESTING IN RELATION TO OBJECTIVES

Any discussion of testing procedures must lack pertinence unless it is related to the objectives of instruction. It is essential, therefore, that before proceeding with our subject we turn our attention to the generally accepted aims of teaching in this particular field.

The most authoritative and widely accepted statement of the aims of mathematical instruction is that contained in *The Reorganization of Mathematics in Secondary Education*, the 1923 Report of the National Committee on Mathematical Requirements.^{*} The reader is referred to that report for the complete statement and discussion of these aims. We outline them here in the following highly condensed summary:

A. Utilitarian Aims

1. Skill in the fundamental processes of arithmetic.
2. Command of the language of algebra.

^{*} *The Reorganization of Mathematics in Secondary Education*, published by the Mathematical Association of America, Inc., 1923.

3. Such knowledge of the fundamental laws of algebra as will equip one to understand and use elementary algebraic methods.
4. Skill in interpreting graphical representation.
5. Familiarity with the geometric forms common in nature, industry, and life. This involves acquaintance with the more fundamental properties and relationships of these forms.

B. Disciplinary Aims

1. The acquisition, in precise form, of those ideas and concepts in terms of which the quantitative thinking of the world is done, and the development of ability to think clearly in terms of such ideas and concepts.
2. The acquisition of mental habits and attitudes which will make the above training effective in the life of the individual: a seeking for relations, an attitude of inquiry, a love of precision, a desire for orderly and logical organization, etc.
3. Training in thinking in terms of the idea of relationship or dependence.

C. Cultural Aims

1. Appreciation of beauty in the geometric forms of nature, art, and industry.
2. Ideals of perfection as to logical structure, precision of statement and of thought, logical reasoning, etc.
3. Appreciation of the power of mathematics.

All these aims are to be interpreted so as to contribute to the general point of view that "the primary purpose of the teaching of mathematics should be to develop those powers of understanding and of analyzing relations of quantity and of space which are necessary to an insight into and control over our environment and to an appreciation of the progress of civilization in its various aspects, and to develop those habits of thought and of action which will make these powers effective in the life of the individual." More specifically, it is urged that "continued emphasis throughout the course must be

placed on the development of ability to grasp and to utilize ideas, processes, and principles in the solution of concrete problems rather than on the acquisition of mere facility or skill in manipulation."

In dealing with the objectives in any field of study, it is frequently helpful to classify them under the general headings "immediate" and "ultimate." Immediate objectives are very directly associated with the actual subject-matter content of courses of study. In fact, to say that a pupil has mastered the subject matter as taught is to say that he has realized the immediate objectives of that teaching. "Understanding of the various methods of proving triangles congruent" will obviously be one of the immediate objectives of instruction in geometry. Ultimate objectives are farther removed from specific subject-matter content. They are superimposed on the immediate objectives and are realized only when the pupil has acquired the power to generalize the knowledge and skills which very largely constitute these immediate objectives. They are, for the most part, in the nature of fundamental beliefs, attitudes, appreciations, ideals, and methods of thinking and reacting in general. For example, an ultimate objective of instruction in geometry is the development of an appreciation of logical organization. Almost any skillful teacher, if working under favorable conditions, will achieve satisfactorily the immediate aims of instruction; only those with the broader vision will aim for and satisfactorily achieve the ultimate objectives.

If we study the aims of mathematical instruction as enumerated above, we shall see, as we should indeed expect, that those in the utilitarian group are largely immediate aims, while those in the disciplinary and cultural groups must properly be considered ultimate aims. Important as these ultimate aims may be, it is obvious that they are highly idealistic and intangible in character. They can be realized only by way of the more

readily attainable immediate aims, and the roads leading from one to the other are seldom clearly defined. The instructional materials now available and in general use are devoted almost exclusively to the development of specific skills and abilities, and current methodology is quite silent with reference to techniques for achieving disciplinary and cultural outcomes through the agencies of such materials. Courses of study may urge on the teacher the importance of ultimate objectives, but they offer him little direction in the manner of their attainment. The teacher has no authoritative guide to point out a workable approach to such problems as those of developing an appreciation of beauty, inculcating ideals of perfection, or establishing a desire for orderly and logical organization: so he contents himself with working toward more tangible immediate goals. He tries to develop in his pupils an understanding of and ability in the application of those mathematical principles and processes which are presented in the text or outlined in the course of study. He trusts that in the very nature of things many disciplinary and cultural values will accrue, but for the most part he is ignorant of techniques which can be counted on to insure their realization. He is disposed to judge the effectiveness of his teaching by the degree to which immediate objectives are attained, and the tests he devises to measure his success are designed accordingly.

The preceding discussion makes it clear that development in teaching and testing has been largely related to the immediate objectives. It is nevertheless very important that success in the attainment of all recognized objectives be evaluated if that can be done successfully. Various experiments are making progress in the direction of evaluating less immediate objectives, such as those related to fundamental processes of reasoning, and those expressed in terms of attitudes and habits. It must be recognized, however, that important as are the ap-

preciations, habits, attitudes, and generalized thinking processes, it is not yet possible to suggest to teachers any generally satisfactory methods of testing progress in establishing them. It is nevertheless important that teachers shall follow the work being done in this field, and shall themselves experiment with methods of evaluating the attainment of what they consider important ultimate aims. So if, throughout this chapter, we seem to pay inadequate heed to the disciplinary and cultural aims of instruction, it must not be assumed that we fail to appreciate their importance or that we condone the lack of emphasis upon them.

TRADITIONAL AND NEW-TYPE TESTS IN MATHEMATICS

In any treatment of testing techniques, one of the main points for discussion will be the relative merits of traditional and new-type examinations. Although this matter has been dealt with in a general fashion in other sections of this manual, it is probably advisable to review the ground briefly with particular reference to measurement in mathematics.

One of the advantages which new-type tests claim to enjoy over traditional examinations is their objectivity. They are scored by use of a key, and if directions are accurately followed the scores are unaffected by subjective factors. This consideration is often of paramount importance. In the case, for instance, of standardized tests designed for use in a wide variety of situations with the results to be interpreted against established norms, objectivity is absolutely essential; and quite as necessary in mathematics as anywhere else. It is often assumed that because of the very nature of the material, the scoring of mathematics papers of any sort must inevitably be more objective than the scoring of papers in most other school

subjects. There is probably some ground for such an assumption if one has in mind the ability of a teacher to estimate the relative merits of a number of test papers, but there seems to be little basis for it if he has reference to the likelihood that a number of different teachers will arrive at close agreement on the marks to be awarded to any particular paper. While the teacher of mathematics may be able to apply, fairly objectively, his own private standards of scoring, these standards probably differ from the standards of other mathematics teachers as widely as they would differ among teachers in any other subject. This was clearly demonstrated in an experiment reported in 1913 by Starch and Elliott. The same geometry paper was scored by 114 geometry teachers. The marks assigned ranged from 28 to 92, with a quarter of the group awarding scores of 62 or under and another quarter granting scores of 78 or over. Such results represent the situation with regard to uniformity of standards fairly accurately, and we are left to conclude that if norms in standardized tests in mathematics are to mean anything they must be interpreted against scores derived objectively.

In informal testing the importance of objectivity depends to a considerable extent on the purposes which the tests are meant to serve. If they are to be used for surveys, whether over wide areas or within single schools, with the scoring to be done by a variety of persons, objectivity is just as imperative as it is in regularly standardized examinations. But when, as is usually the case, the informal test is designed for use by a single teacher, objectivity loses a good deal of its importance if the teacher is guided, in his scoring, by a reasonably reliable system from which he is not easily swayed by irrelevant considerations. Such systems are probably devised more readily in mathematics than in most other subjects. The science of mathematics is itself perfectly objective. Answers tend to be

definitely right or wrong, with little room for any middle ground. Processes are carried forward, as a rule, along conventional lines, and such extraneous factors as literary style and conformity to grammatical usages are not likely to becloud the issue. But while, for these and other reasons, the teacher of mathematics has fewer obstacles to overcome than have many others in his efforts to effect reliable scoring, we must not assume that he can therefore afford to scorn the virtues of techniques which by their very nature are completely objective.

A second major advantage of the new-type tests is the convenience with which they may be used to break up the testing material into small units. There is a three-fold virtue in this. In the first place, it makes possible a wider sampling of content, since many more problems can be worked in a given unit of time. In the second place, it affords an opportunity of ascertaining the pupil's understanding of the crux of a problem situation without burdening him with needless computation and manipulation. Finally, it offers better diagnostic possibilities. When a pupil fails in the solution of a long and involved problem, it is often difficult to locate his trouble; but when the examination is presented in small units, the matter of diagnosis is usually facilitated.

New-type tests enjoy a number of additional virtues. They are unsurpassed for certain forms of drill. For the most part, students prefer them to the traditional types of examination. They exist in so many forms that they are unusually adaptable to a wide variety of situations. They hold the student to the particular issue and focus his thinking in the desired directions. They tend to make the teachers who handle them increasingly conscious of the importance of scientific examination techniques, with a resulting elevation of standards of test construction, administration, and interpretation. A final feature,

and one which has done much to popularize new-type tests, is the time-saving they effect in scoring. In subjects such as history or English this may be of considerable importance, especially where classes are large and examinations frequent. It does not, however, constitute a very compelling argument for their use in mathematics, for here the scoring of papers has always been a comparatively light burden. In fact, if the proper care is bestowed on the preparation of new-type tests, the time thus spent by the teacher of mathematics will often be greater than the time saved in evaluating the papers.

It must not be assumed from the foregoing statements that new-type tests are adequate for all purposes. They have their weaknesses and, as far as their everyday use is concerned, they should supplement rather than supplant traditional examinations. Their general use is somewhat limited in several fundamental areas of achievement in mathematics, and there are certain types of objective techniques which have quite serious limitations in the measurement of achievement over specific content or of ability in specific skills.

There are certain purposes for which questions of the essay type are definitely superior to any that have yet been devised in the new objectively scored forms. Among these purposes are those of testing for the organization of information in terms of the student's own generalizations; testing for the student's abilities to make use of the knowledge he has acquired in relatively complex, sustained reasoning situations; and testing for his ability to rationalize the procedures he has employed. It is not feasible to attempt to present and discuss here all of the types of essay questions which may thus be used to advantage. The following examples, however, are indicative of what the teacher can do in testing for outcomes that are not adaptable to objective test forms now in use.

EXAMINATIONS IN MATHEMATICS

1. Make as general a statement as you can concerning the conditions necessary to prove two triangles congruent.
2. Under what conditions is a quadrilateral known to be a parallelogram?
3. What ways do you now know by which two angles can be proved equal? unequal?
4. Which proposition studied this (week, month, quarter, as preferred) seems to you most difficult to prove? Explain why you find it difficult.
5. Show the relation between congruence of triangles and similarity of triangles in terms of the conditions required.
6. How much of a circle is needed to find its center? Why?
7. Give the derivation of "geometry," and explain some need that forced an early civilization to discover the rudiments of this science.
8. Write a short account of the contribution of Euclid to geometry.
9. What limitations keep Euclidian geometry from more completely covering a wider range of plane figures in its constructions?
10. What other statements are closely related to "If A is true, B is true"? Which of them follow from this given one?
11. If you have proved the theorem "The base angles of an isosceles triangle are equal," can you conclude that if the base angles of a triangle are equal, the triangle is isosceles? Explain your answer completely.

SUGGESTIONS FOR TEST CONSTRUCTION

The general and specific uses of new-type tests will be discussed in the following pages. The examples used are drawn only from algebra and plane geometry but should apply with equal force to any subject in the curriculum of secondary mathematics.

Suggestion A

As far as possible, test items should require the application of generalized ideas, processes, and principles rather than the

mere ability to recall verbal statements of relationships, formal rules, or definitions. For the most part, the problem should be presented in a form that differs from the conventional textbook presentation which students often master in a more or less mechanical fashion.

There is little difficulty in satisfying this principle in algebra as far as verbal problem material is concerned. Almost any verbal problem makes reasonably satisfactory demands on generalized ability to handle the processes involved. It is in the simpler, more routine procedures that we must be on our guard. The following examples illustrate various ways in which test items may call for the same very simple operation:

12. Divide $x^2 - 7x + 12$ by $x - 3$
13. $x - 3$ is one of the two factors of $x^2 - 7x + 12$. What is the other factor?
14. By what must $x - 3$ be multiplied in order to obtain $x^2 - 7x + 12$ as a product?
15. What is the quotient obtained by using $x - 3$ as a divisor and $x^2 - 7x + 12$ as a dividend?
16. If the area of a rectangle is represented by $x^2 - 7x + 12$, and the length by $x - 3$, what expression will represent the width?

The first of these items tests *directly* and formally the pupil's ability in simple algebraic division. The other items, while likewise testing this ability, further demand that the pupil be able to recognize that division rather than some other process is called for, and also that he have acquired certain other mathematical information of a general nature. Items of this latter type place a distinct premium upon generalized ability to apply the skills and processes presumably mastered by the pupil.

The available instructional materials in algebra are seriously deficient in the emphasis given to development of the type of generalized ability necessary for success in the last four of the

EXAMINATIONS IN MATHEMATICS

above items. For this reason the following pairs of test items are presented to illustrate further the important distinction between "formal" and "functional" testing of achievement in this area of mathematics. The first item (a) of each pair is stated in the conventional textbook form, while the second (b), though usually involving the same mathematical process, is rephrased so as to require some generalization or transfer in the application of that process.

17. (a) Solve for x : $\frac{5(x+3)}{2} = 0$

(b) For what value of x will the expression $\frac{5(x+3)}{2}$ be equal to zero?

18. (a) Solve for n : $n^2 - 1 = 0$

(b) For what two values of n will $\frac{1}{n}$ be equal to n itself?

19. (a) Graph the equation $3x - 2y = 6$

(b) If the graph of the equation $3x - 2y = k$ passes through the point $(4, 3)$, what is the value of k ?

20. (a) Solve $3n^2 + 5n - 2 = 0$

(b) For what fractional value of n will the expression $3n^2 + 5n - 2$ be equal to zero?

21. (a) Solve $3x - 7 = 8$

(b) In the equation $3x - k = 8$, what must be the value of the constant k if 5 is the root of the equation?

22. (a) Simplify $\frac{4a}{2x-y} + \frac{3a}{2y-4x}$

(b) The fractions $\frac{4a}{2x-y}$ and $\frac{3a}{2y-4x}$ may be added together and the result simplified to form a single fraction with $2(2x-y)$ as denominator. What is the numerator of this fraction?

23. (a) Solve for x : $\frac{a+x}{b} = \frac{c+x}{d}$

- (b) What quantity must be added to the numerators of the fractions $\frac{a}{b}$ and $\frac{c}{d}$, respectively, in order to make their values equal?

In geometry there seems to be little justification for isolated testing of ability to recall formal statements of propositions, axioms, rules, principles, or definitions. Knowledge of such matters is useful only as the student appreciates and understands them in their applications, and tests should be formed accordingly. Fortunately, this type of functional testing is not difficult to achieve in the field of geometry. Almost any "original" geometric exercise of the traditional sort is functional in nature. The following items illustrate how the same characteristics can be achieved for material in an objective form.

24. The angles of a triangle are in the ratio 2 to 3 to 4. What is the size, in degrees, of the largest angle?
25. What is the number of degrees in an angle that is one-fifth as large as its complement?
26. A circle can be circumscribed about any (1) rhombus, (2) equilateral hexagon, (3) parallelogram, (4) equiangular pentagon, (5) isosceles trapezoid. ()
27. Which is the shortest side of the triangle ABC if $\angle A > \angle B > \angle C$?

In agreement with the principle set forth in this discussion, the algebraic notation used in the problem situations should in many instances differ from the usual x, y, z and a, b, c notations of the textbooks. The difficulties experienced by students in high-school physics demonstrate clearly the necessity for more highly generalized ability in the use of mathematical symbols. Many students are so accustomed to thinking of equations in

EXAMINATIONS IN MATHEMATICS

terms of x , y , and z that they become confused when they meet them in any other dress. Items 28 to 31 are identical in method of solution, yet almost any class will find them varying quite widely in difficulty. The obvious cure for such a situation is plenty of practice with a variety of notations.

28. If $3 = \frac{2}{x}$ then x equals..... ()

29. If $a = \frac{b}{x}$ then x equals..... ()

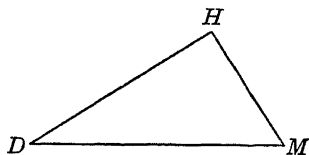
30. If $D = \frac{W}{F}$ then F equals ()

31. If $e = \frac{5}{K}$ then K equals ()

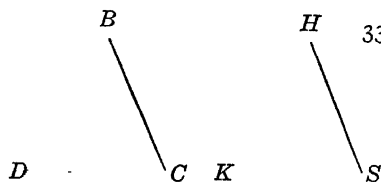
Similarly with geometry. Probably it is even more necessary here than in algebra that we depart from the exact figures and notations used in the textbooks. A pupil will hardly be guilty of following so closely the text in algebra as to memorize, verbatim, the solutions for specific problems, but in geometry this practice is quite prevalent. Pupils whose grasp of geometric processes is very feeble will commit to memory, in complete detail, the proofs of the assigned propositions.

Tests should be designed to discriminate against those who rely on such practices. The illustrative figures and diagrams used should not be exact reproductions of textbook figures but should require generalized ability to recognize basic relationships regardless of the external characteristics of the figures employed. Practically all teachers of mathematics have had experience with pupils who do not recognize two parallel lines as such unless the lines are drawn horizontally or vertically on the blackboard or page. Other pupils may have failed to realize that a right triangle does not necessarily need to be drawn with the right angle at the lower left-hand corner of the figure.

Pupils with such an inadequate understanding may quite readily respond correctly to a test item involving a figure in text-book form, but fail when some variation is introduced. In fact, if the ability to demonstrate a theorem is wholly a product of memorization the mere matter of altering the lettering of a diagram may be sufficient to cause failure. If, in addition to altering the lettering, changes are introduced in the positions and proportions of the diagrams, the difficulty of the problem is still further increased. Many pupils who are able to demonstrate correctly that "the base angles of an isosceles triangle are equal," or that "two triangles are congruent if the three sides of one are respectively equal to the three sides of the other," will fail completely when the problems are presented in the following manners:



32. In $\triangle HDM$, $HD = MD$.
Prove $\angle H = \angle M$.



33. In $\triangle BDC$ and HKS , $BD = HK$, $BC = HS$, $DC = KS$.

By placing $\triangle HKS$ so that HK coincides with BD , prove $\triangle HKS \cong \triangle BDC$.

Suggestion B

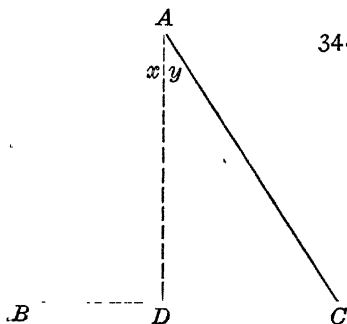
Some of the more fundamental processes in mathematics do not lend themselves well to measurement by new-type techniques.

This observation applies with particular force in the field of geometry. One of the most important immediate aims in geometric instruction is the development of ability to work one's way by logical sequences of arguments to the proofs of geometric propositions. This ability is fostered by practice in

demonstrating the theorems prescribed in courses of study and in applying these theorems in solving originals.

The measurement of achievement in logical demonstration presents two major considerations to test makers. On one hand the traditional exercise, in which the pupil is asked to present a complete logical proof of a proposition, involves subjective scoring, the unreliability of which has repeatedly been demonstrated. Adequate scoring of such exercises, even by the individual teacher for his own group of pupils, is a task of great difficulty and a practical impossibility for testing programs of school- or state-wide extent. On the other hand, the new-type tests thus far devised, particularly those based on recognition rather than recall, do not measure adequately the ability under consideration. In presenting various new-type devices which test makers have employed, we do so with no intention of setting up these techniques as models of good practice. It is rather our wish to present this situation in its entirety to teachers and test makers alike in the hope that serious attention to the problem involved will lead to further progress in the field of achievement testing.

In the first of these devices the steps of the proof are presented in disarranged order. It is the task of the testee to indicate their proper sequence.



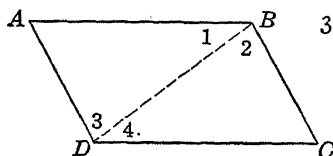
34. Given $\angle ABC$, with $AB = AC$.
To prove $\angle B = \angle C$.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

Statements (without reasons) in jumbled order	Correct order of statements
1. $AD = AD$	() comes first
2. $\angle B = \angle C$	() comes second
3. $\angle x = \angle y$	() comes third
4. Bisect $\angle A$ by AD	() comes fourth
5. $\triangle ABC \cong \triangle ACD$	() comes fifth
6. $AB = AC$	() comes sixth

A very obvious fault of such a set-up is the fact that it is almost impossible to find a theorem in which the steps in the proof follow one another in unique sequence. Even in the very simple example above there are half a dozen or more different arrangements which are equally valid, and in longer problems the confusion would likely be so great as to render scoring keys hopelessly complex. A more serious fault is the fact that such a method measures very superficially the ability it should be trying to test. Indicating the correct sequence of presented steps in a proof is quite a different matter from the really fundamental task of developing and reproducing the proof unaided by any cues.

Another device sometimes employed in published tests consists of presenting the demonstration of the theorem with certain steps omitted. The student's ability is estimated on the basis of his skill in filling in the blanks correctly.



35. Given parallelogram $ABCD$

To prove $AB = DC$

Proof (reasons omitted)

1. Draw DB

2. _____

3. $\therefore \angle 1 = \angle 4$

4. $AD \parallel BC$

5. \therefore _____

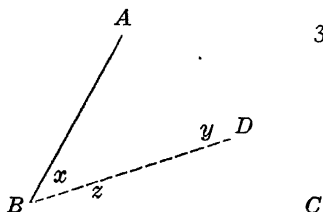
6. _____

7. $\therefore \triangle ABD \cong \triangle CDB$

8. \therefore _____

This device is an improvement over the one previously presented in that the responses are definitely located. The problem can always be so arranged that the statement to be entered in each blank is rigidly prescribed by the nature of the immediate context. However, this type of item can hardly escape the major criticism that as a means of measuring ability to demonstrate theorems it is a poor substitute for the real thing.

A third device is that of presenting the proof with one unnecessary or wrong step. The student must locate the error.



36. Given $\triangle ABC$, with $AC > AB$.
To prove $\angle B > \angle C$.

Which step in the following proof is wrong or unnecessary?
(reasons omitted) ()

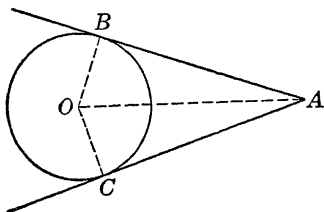
Proof

1. On AC lay off $AD = AB$
2. Draw BD
3. $\angle ABC > \angle x$
4. $\angle x = \angle y$
5. $\angle y > \angle C$
6. $\angle x > \angle z$
7. $\angle ABC > \angle x > \angle C$
8. $\therefore \angle ABC > \angle C$

This device has the virtue of perfect objectivity, but, like the others considered, it yields no very valid measure of the ability it is attempting to evaluate. Ability to discover discrepancies in a presented proof is surely a very superficial indication of ability actually to develop the proof. Moreover, the method is somewhat ponderous, in view of the fact that it

yields but a single response. And if one were to try to remedy this defect by increasing the number of errors, the matter would easily become so complicated as to be quite useless.

A fourth arrangement, and the last we shall consider, is one in which the proof is presented with omissions in both statements and reasons. The student completes the set of statements by filling in the blanks. An appended list of reasons allows him to indicate the correct missing reasons by number.



37. Given circle O with tangents AB and AC .

To prove $AB = AC$

<i>Statements</i>	<i>Reasons</i>
1. Draw OB , OC , and OA	1. (3)
2. $\angle OBA$ and $\angle OCA$ are rt \angle 's	2. ()
3. _____	3. (8)
4. _____	4. (1)
5. $\therefore \triangle OBA \cong \triangle OCA$	5. ()
6. _____	6. ()

Reasons

1. Identity.
2. Two sides and included angle = two sides and included angle.
3. A straight line can be drawn between two points.
4. Corresponding parts of congruent triangles are equal.
5. An angle inscribed in a semicircle is a right angle.
6. The tangent to a circle is perpendicular to the radius drawn to the point of contact.
7. Hypotenuse and leg = hypotenuse and leg.
8. Radii of the same circle are equal.
9. An angle formed by a tangent to a circle and a chord to the point of contact is measured by half the intercepted arc.

This seems to be a more valid measuring device than any of the three preceding ones, although its advantage in this respect is diminished by the fact that it is gained at the cost of considerable complication in the machinery of the problem. The matching process involved is also open to criticism in that the statements given as "Reasons" do not have the homogeneity that is desirable.

It is evident that none of these devices is completely satisfactory as a measuring instrument for ability in logical demonstration. Ideally, this ability should entail a grasp of geometric truths and processes that will enable one to carry through the required work of analysis and organization independently of aid. This is particularly true in the case of originals, where, by the very nature of things, originality and initiative are all-important. The very essence of ability to deal with originals is ability to "see through" the situation, ability to initiate and carry through, unaided, the necessary sequence of arguments. Devices such as those indicated here can hardly be deemed to measure this ability with any real effectiveness. When used in standardized tests they may be regarded as concessions to the necessity for complete objectivity, and they will hardly appeal to the teacher for use in his informal classroom examinations. As far as ability in demonstrating geometric propositions is concerned, he will be well advised to do his measuring through the medium of traditional examinations scored as objectively as possible.

It is important, however, to remember that all geometric problems involving a necessity for logical reasoning, originality, and initiative need not necessarily fall into the conventional classification of "proofs of geometric propositions." Test items of the short-answer type may readily be constructed which cannot be solved by the pupil unless his reasoning is

logical and sustained, even though he is not asked to present an organized demonstration of proof. Such items may also necessitate a quite comprehensive understanding of basic geometric principles and processes. The following examples illustrate the truth of these assertions:

38. Given an isosceles triangle one of whose equal sides is 6. What is the perimeter of the parallelogram formed by drawing parallels to the equal sides from any point P on the base? ()
39. In the acute triangle ABC , $\angle A > \angle B > \angle C$. E , F and G are the midpoints of sides AC , AB , and BC respectively. What is the shortest side of triangle EFG ? ()
40. From an external point P a tangent PA is drawn to a circle with center at O . PO cuts the circle at B , and AC is a perpendicular from A to PO . How many degrees are there in angle PAC if angle PAB is 25 degrees? ()
41. A triangle ABC is inscribed in a circle. The bisector of angle BAC intersects BC at P and cuts the circle at D . What is the length of BD if AP is 8 and PD is 2? ()

Items such as these are admittedly time-consuming but not nearly so much as the traditional exercise used in proving geometric propositions. They have also the virtue of complete objectivity and furnish a fairly satisfactory approach to the measurement of ability in geometric reasoning.

The situation in another very important area in geometry is somewhat similar. There has been developed no wholly satisfactory method of measuring ability to make accurate geometric constructions. It is not that objectivity is altogether impossible, but rather that it is achieved at too great a cost. The student, for instance, might be given some such problem

as the following, with credit awarded on the basis of the accuracy of the measurement recorded:

42. Construct an equilateral triangle with sides 6 centimeters long. Construct an altitude. What is its length in millimeters? ()

This type of problem is frequently employed, and we cannot deny that it is perfectly objective. However, in regarding only the accuracy of the end-result, it ignores one fundamental element in the situation, that is, the process by which this end-result is achieved. It may be argued that an accurate answer is sufficient guarantee of accuracy in the contributory steps. Such a position, however, is hardly tenable. In many construction problems the student can derive acceptably accurate answers by rule-of-thumb methods which completely ignore geometric considerations. In Item 42, for example, with the triangle properly constructed, it is not necessary that the altitude be drawn by the usual processes at all. Almost any line drawn from a vertex to the opposite side in such a manner that it "appears to be perpendicular" will be accurate enough in length to pass muster. The extent to which pupils resort to such rule-of-thumb methods is, of course, problematical but must be considered in constructing items designed to measure directly the pupil's ability to make *accurate* geometrical constructions.

It seems clear, however, that an extensive use of construction problems in general achievement examinations may be readily justified. The inclusion of objective construction exercises in an achievement test in plane geometry has as its function far more than the testing of ability in the unique mechanical skills involved in the actual process of constructing an accurate figure. The original construction of geometric figures requires more of the pupil than mere ability to recognize re-

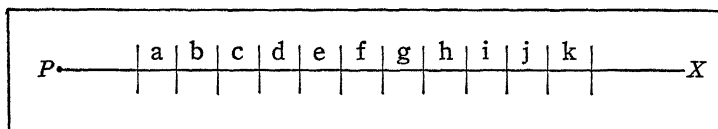
EXAMINATIONS IN MAJOR SUBJECT FIELDS

relationships or to recall facts and principles which may have been learned by rote. It calls for the *application* of these facts and principles in new and original situations, situations which are particularly effective because of their non-verbal character. It therefore places a distinct premium upon a reasoned understanding of geometrical facts and processes and tends to negate the effects of verbalism in learning. In general, test exercises requiring constructions should not be considered as only measures of skill with compasses and straightedge, but rather as comprehensive measures of nearly all aspects of geometry, since they may call into play many of the skills, abilities, or information that may have been acquired throughout the course of instruction.

The following type of construction exercise has been found to be useful in measuring achievement in plane geometry.

Directions: In each exercise follow the directions carefully, performing each step in the order given. In every case your answer will consist simply of one letter to be written in the blank provided.

Use this scale as directed in each exercise.



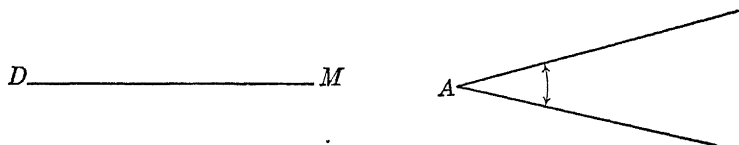
43. (a) Construct a triangle ABC with angle A equal to 45° and angle C equal to 75° . The base AB of the triangle is given below.

A _____ B

- (b) Measure the side AC with your compasses and then lay it off on the scale PX . To do this, use point P on the scale as a center, take the length of AC as a radius, and draw an arc cutting the scale PX somewhere between P and X .
- (c) Which of the segments $a, b, c, \dots k$, of the scale PX , is cut by the arc? Answer _____

EXAMINATIONS IN MATHEMATICS

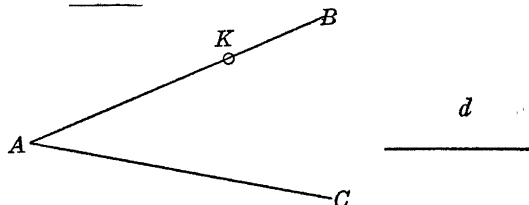
44. (a) Inscribe the given angle A in a circle whose diameter is the segment DM .



- (b) Measure with your compasses the chord of the arc intercepted by angle A and then lay it off on the scale PX .
 (c) Which of the segments $a, b, c, \dots k$, of the scale PX , is cut by the arc?

Answer _____

45. (a) Locate two points, Q and R , equidistant from the sides of angle BAC and also at a distance d from a given point K .



- (b) Measure the distance QR with your compasses and then lay it off on the scale PX .
 (c) Which of the segments $a, b, c, \dots k$, of the scale PX , is cut by the arc?

Answer _____

Pupils taking such a test should be instructed to provide themselves with only a pair of compasses and a straightedge, the latter preferably unruled. The test exercises may be so spaced that the actual constructions can be performed on the test page itself. Each exercise should be so devised that, if the pupil's work is completely accurate and the required line-

EXAMINATIONS IN MAJOR SUBJECT FIELDS

segment is laid off along the scale PX , the arc drawn will fall at the exact midpoint of one of the scale intervals. If the measured line-segment is not completely accurate but is within one-half of an interval of the correct length, the arc will fall within the proper interval, and the pupil will be credited with a correct response. Allowance is thus made for the slight inaccuracies which may enter into construction when compasses and straightedge of inferior quality are used. The margin allowed for errors may, of course, be adjusted by varying the lengths of the segments $a, b, c, \dots k$ on the scale PX .

The problem of evaluating ability in solving verbal problems in algebra is also perplexing. Of course the traditional method is itself highly objective, and the more simple problems, or those involving very few steps in their solutions, are probably best dealt with in this manner as illustrated by the following examples.

46. A dealer sold a suit for \$60, making a profit of 25 per cent of its cost. How many dollars profit did he make? ()
47. The area of a triangle is 36 square inches, and the altitude is 9 inches. What is the length of the base? ()

For the more difficult verbal problems, however, this type of item tends to be too time-consuming and involves a unit which is often too large to suit the general character of the new-type tests. The task for the examiner, therefore, resolves itself into finding techniques which test *some one crucial aspect* of a problem situation without forcing the pupil to perform a number of time-consuming or otherwise routine steps in its solution. The following example illustrates an attempt to isolate one crucial aspect of a problem situation, that of forming the necessary equation, and to reduce the size of the unit by assisting the student over a number of routine steps.

48. The length of a rectangle is three times its width. A second rectangle is 2 inches wider and 3 inches shorter than the first and has an area 6 square inches greater. What equation would you use to find the width of the first rectangle?

Let x = the width of the first rectangle

then $3x$ = the length of the first rectangle

$x + 2$ = the width of the second rectangle

$3x - 3$ = the length of the second rectangle

$3x^2$ = the area of the first rectangle

$3x^2 + 3x - 6$ = the area of the second rectangle

I would use the following equation (_____)

Such a device has a number of faults. In the first place, the machinery employed is rather cumbersome in view of the fact that only a single response is involved. A more serious consideration lies in the fact that the particular unit of work left for the student to carry out may not be a very valid measure of the ability it is attempting to evaluate. It should be noted that the steps which may be considered as routine will vary in character from problem to problem and from one group of pupils to another. Setting up the final equation may sometimes be quite as routine a matter as that of giving correct algebraic expression to the contributory steps. Ability in problem-solving consists very largely of the initiative and insight required in setting up the sequence of steps leading to the desired equation; in freeing the student from the necessity of setting up these steps for himself, one of the significant parts of the problem is eliminated. The use of this device also leads to scoring difficulties, as the equations required may often be capable of a great variety of forms.

The following example illustrates an alternate form of the above technique.

49. The length of a rectangle is three times its width. A second rectangle is 2 inches wider and 3 inches shorter than the first

EXAMINATIONS IN MAJOR SUBJECT FIELDS

and has an area 6 square inches greater. What is the width, in inches, of the first rectangle?

Let x = the width, in inches, of the first rectangle.

Then:

- (a) The length, in inches, of the first rectangle is, in terms of x ()
- (b) The area, in square inches, of the first rectangle is, in terms of x ()
- (c) The width, in inches, of the second rectangle is, in terms of x ()
- (d) The length, in inches, of the second rectangle is, in terms of x ()
- (e) The area, in square inches, of the second rectangle is, in terms of x ()
- (f) I would use the following equation to find x (solution not required). ()

This device also is open to criticism on the ground that it suggests to the student the proper sequence of steps to be followed. Moreover, it leaves to the pupil the routine matter of supplying a number of simple algebraic representations, thereby making the item needlessly cumbersome in view of its avowed purpose. Ability in making correct algebraic representations may be tested in other single items, as may also the ability to solve equations. The really significant ability demanded by verbal problems, aside from the initiative and insight required to formulate methods of approach, is that of setting up the necessary equations.

In trigonometry we find a similar situation. Here the whole course of study converges on the solution of triangles, and no new-type techniques have as yet been devised as substitutes for traditional methods of measuring ability in this area. From the very nature of things, we could hardly expect the situation to be otherwise. As is the case with demonstrating theorems in geometry and solving the more difficult verbal problems in

algebra, the matter of solving triangles in trigonometry is a rather complex process. It involves a long and logical sequence of steps and constitutes a larger unit than new-type techniques are well designed to handle.

Suggestion C

Among objective types of items the short-answer is, in general, the most suitable for use in mathematics.

As compared with the multiple-choice and other recognition forms of objective items, the short-answer form enjoys a number of points of superiority.

1. It almost completely eliminates unreliability due to guessing and thus avoids one of the most common criticisms leveled at objective tests.

2. It gives reasonable assurance that the student will arrive at the answers to the questions by the route intended. Often, in recognition forms, if the items have not been constructed with care, the correct responses can be identified by merely checking the alternatives against the data presented. Flagrant instances of this fault are seen in the following examples.

50. The square root of 544644 is (1) 748 (2) 742 (3) 738
 (4) 732 (5) 728..... ()
51. Factor: $6x^2 - 7x - 5$ (1) $(6x + 5)$ $(x - 1)$
 (2) $(2x - 5)$ $(3x + 1)$ (3) $(x - 5)$ $(6x + 1)$
 (4) $(3x - 5)$ $(2x + 1)$ (5) $(2x + 5)$ $(3x - 1)$ ()
52. The positive root of the equation $x^2 - 7x - 18 = 0$ is
 (1) 1 (2) 2 (3) 3 (4) 6 (5) 9..... ()

Item 50 obviously is intended to measure the student's ability to carry through the process of extracting the square root of an arithmetical quantity. Yet the examinee who knows no more about square root than the mere definition of the term will, if he possesses even a modicum of intelligence, change the problem into one of simple multiplication. He will square each

EXAMINATIONS IN MAJOR SUBJECT FIELDS

of the responses in turn until he arrives at the one that yields 544644 as a product. And similarly with the other two. In No. 51, the student can again replace the intended process by one of multiplication and arrive at the solution quite as surely, even though not as directly. All he needs to know in order to identify the right response in Item 52 is the technique of substituting in an equation to check for the correctness of a root.

The improvement resulting from the use of the short-answer form for items such as these is obvious.

- 53. What is the square root of 544644?..... ()
- 54. Write $6x^2 - 7x - 5$ as the product of two factors..... ()
- 55. What is the positive root of the equation $x^2 - 7x - 18 = 0$?..... ()

3. The simple recall form is economical of space on the test paper. A superficial examination of the items used as illustrative material throughout this discussion will show that short-answer items are less than half as space-consuming as are multiple choice. This may be an important consideration when facilities for reproducing individual test papers are limited.

4. For situations in which it is appropriate, it is the easiest of all objective forms in which to frame test material. Probably less skill and less training are required for the construction of short-answer items than for any other new-type form. It is the sort of thing with which all are acquainted. We go through life asking and answering such questions as: "What time is it?" "How much did it cost?" "Were you at the dance?" When one constructs short-answer items, he is simply doing, with somewhat greater attention and care, the sort of thing he is doing continually as part of the daily routine of living.

- 5. Whether framed as a question or as an incomplete state-

ment, the short-answer form presents the problem to the student in a setting with which he is wholly familiar. While this is a distinct advantage in tests for use with very young children or those unfamiliar with new-type techniques, it decreases in importance as they grow older and become test-wise.

If we turn now to the reverse side of the picture we shall see that the short-answer form suffers at least three minor disadvantages.

1. Ordinarily, short-answer items are less readily scored than are the various recognition types. Many of the answers will be somewhat more complex and will require closer scrutiny than will the single digit type of answer characteristic of matching and multiple-choice items. Moreover, it is not always possible to cast test material in the short-answer mould with assurance that each item admits of one and only one form of answer. Consider the following:

56. What is the quotient when $-10a^5b^{10}$ is divided by $-2a^5b^5$?..... ()
57. What is the square root of $9a^2 - 12ab^2 + 4b^4$?.... ()
58. If $a - cx = n$, what does x equal?..... ()

If we study Items 56 and 57, we see that there is little likelihood that the correct answer to either will be expressed in more than one form. For Item 58, however, we might expect the answer in any of three variations: $\frac{a-n}{c}$, $\frac{n-a}{-c}$, $-\frac{n-a}{c}$. Such

a circumstance makes it necessary either to complicate the scoring key by the inclusion of alternative answers, or to have the tests scored only by people so well grounded in mathematics that they will recognize the answer in its various forms.

This fault in the short-answer type of item is of little or no importance except in the case of standardized tests which may sometimes be scored in large numbers by clerical workers

whose limited knowledge of the field makes them wholly dependent on the scoring key. It need hardly be considered a fault at all if we think of it with reference to informal tests administered and scored by a classroom teacher.

2. Unless the responses are quite brief, the task of writing them down may require a considerable portion of the student's time. This, again, is hardly a serious criticism. Responses to short-answer items must, by definition, never be long, and the time spent in recording them will likely be no more than the time saved in reading and grasping items in this simpler form. If a student covers multiple-choice items more rapidly than he does short-answer items of comparable difficulty, there is a strong probability that no genuine saving is involved. His greater speed with multiple-choice material is more likely attributable to the fact that he is either guessing or responding to superficial clues in many instances instead of arriving at the answers by the intended processes.

3. In the construction of an achievement test consisting of short-answer items, care must be exercised to prevent over-emphasis on items that are purely factual in nature or measure only elementary skills and items of information.

It should be recognized, however, that a number of such items must be included in any achievement test if the test is to discriminate between pupils at the lower levels of achievement. In fact, although the tests prepared for widespread use are often criticized as being too factual in nature, analysis of test results provides evidence that test makers even now tend to over-rate considerably the achievement level of pupils in secondary mathematics. It is essential, however, to include some items, in short-answer form, which necessitate an appreciative grasp of fundamental mathematical principles and processes. As illustrations of such items the following examples are presented.

EXAMINATIONS IN MATHEMATICS

59. The area of a certain square is represented by the expression $9x^2 + 6xy + y^2$. What is the perimeter P of this square, expressed in terms of x and y ? $P =$ _____
60. Given two equations $x + y = 6$ and $x - y = 2$. What will be the x -value of the point in which the graphs of these two equations intersect? _____
61. The diameter AB of a given circle is 14 inches. What is the greatest possible area of a triangle whose base is the diameter AB and whose vertex angle C is on the circumference of the given circle? _____
62. The bases of a trapezoid are 3 inches and 6 inches, respectively, and each leg is 3 inches. How many degrees are there in the angle formed by a diagonal and either base? _____

Analysis will reveal that each of these items requires of the pupil a genuine understanding of some general principle or principles, in addition to a knowledge of associated mathematical skills and informations. The short-answer item, then, need not necessarily be confined to the measurement of mere skills or knowledge of isolated facts, but may serve a much broader purpose in achievement testing.

If we review this discussion, we see that the advantages of the short-answer form in contrast with the multiple-choice form of item far outweigh its disadvantages. Some of the advantages are quite fundamental; the disadvantages are somewhat superficial. In stating the case as we have, we may have made it appear stronger than circumstances warrant. It must not be assumed that there is no place for multiple-choice items in mathematics tests. Quite the reverse, as we shall now attempt to show.

Suggestion D

Multiple-choice items may be used advantageously when the correct response to a given problem is somewhat involved or can be written in several correct forms.

Examples 71 and 72 illustrate this use when the responses are fractional or irrational in nature.

Multiple-choice items also offer a technique for measurement of specific skills or specific content which otherwise are difficult to approach. The following examples present illustrations of such usage.

For locus problems in geometry.

63. The locus of the midpoint of a chord of fixed length in a given circle is
- (1) two diameters intersecting at right angles
 - (2) four points equidistant from the center of the given circle
 - (3) the diameter perpendicular to the chord at its midpoint
 - (4) a circle having the same center as the given circle
 - (5) two intersecting arcs within the given circle

Answer _____

64. What is the locus of the centers of all circles passing through two fixed points A and B ?
- (1) The line AB , joining the two points
 - (2) A circle with line AB as its diameter
 - (3) Two points on the perpendicular bisector of line AB
 - (4) Two lines through points A and B , perpendicular to the line AB
 - (5) A line perpendicular to line AB at its midpoint

Answer _____

For problems in dependence and relationships in algebra.

65. In the expression $A = \frac{BC}{D-E}$, the letters A, B, C, D , and E

represent quantities that are always positive. Which one of the following operations will be *sure* to decrease the value of A ?

- (1) Let B and C be constant, and decrease D and E
- (2) Let C and E be constant, and decrease B and D
- (3) Let B and E be constant, and increase C and D
- (4) Let B and D be constant, and increase C and E
- (5) Let D and E be constant, and decrease B and C

Answer _____

66. Which one of the following relations must exist between x and y if the value of the fraction $\frac{x+n}{y+n}$ is to remain constant for any given value of n ? (1) $x > y$ (2) $x = n \cdot y$ (3) $x = y$
 (4) $x = \frac{1}{n} \cdot y$ (5) $x < y$

Answer _____

For problems involving generalizations.

67. If, from a point P , a tangent is drawn to each of two concentric circles, which of the following is true?
 (1) The tangent to the larger circle will be the longer
 (2) The tangent to the smaller circle will be the longer
 (3) The tangents will be equal in length
 (4) It is impossible to say which tangent will be the longer

Answer _____

For new approaches to conventional ideas.

68. Let C and D be the measured circumference and diameter, respectively, of a given circle. Which one of the following expressions indicates mathematically what should be done with C and D in finding an approximate value of π ?

- (1) $\frac{2D}{C}$ (2) $\frac{D}{C}$ (3) CD (4) $\frac{C}{D}$ (5) $\frac{2C}{D}$

Answer _____

69. The graph of the equation $2x - 3y = 0$ may be described as

- (1) a line parallel to the x -axis
- (2) a line parallel to the y -axis
- (3) a single point — the origin itself
- (4) a line passing through the origin and lying in the first and third quadrants
- (5) a line passing through the origin and lying in the second and fourth quadrants

Answer _____

The above examples do not present an exhaustive list of applications for the multiple-choice technique. Additional applications will certainly occur to the ingenious teacher or test constructor. A caution should be expressed, however, against forcing mathematical content into the multiple-choice form when achievement over such content can readily be measured by items in the short-answer form.

Suggestion E

When multiple-choice items are employed, a deliberate effort should be made to introduce incorrect responses which are as plausible as possible to pupils deficient in the ability which the items are designed to measure. The purpose should be so to present incorrect responses that they will tend to be selected in preference to the correct response by pupils who respond on a superficial basis.

In multiple-choice items, there is a tendency for pupils to select the answer on the basis of only a superficial recognition of its external characteristics. They often can see that a certain response "looks better" than the others, even though they do not understand thoroughly the principles involved. It is impossible to prevent pupils from answering multiple-choice items on this basis, but it is possible to make the wrong responses "look better" than the correct response, and so penal-

EXAMINATIONS IN MATHEMATICS

ize the student who depends upon superficial clues. The following items illustrate this argument.

70. The radii of two given circles are R and r , respectively. The area of the first circle is four times that of the second. Which one of the following equations expresses the relationship between the two radii?

(1) $R = 16r$ (2) $R = 4r$

(3) $R = \frac{r}{2}$ (4) $R = 2r$ (5) $R = \frac{r}{4}$ ()

71. If $\frac{1}{R} = \frac{2}{b} + \frac{5}{c}$, then R equals (1) $\frac{b+c}{7}$ (2) $\frac{b}{2} + \frac{c}{5} - 1$

(3) $\frac{bc}{2c+5b}$ (4) $\frac{b}{2} + \frac{c}{5}$ (5) $\frac{bc}{2b+5c}$ ()

72. The product of $\sqrt[3]{a}$ and $\sqrt[3]{a}$ is (1) $\sqrt[6]{a}$ (2) $\sqrt[5]{a^2}$

(3) $\sqrt[6]{a^5}$ (4) $\sqrt[5]{a}$ (5) $\sqrt[6]{a^2}$ ()

In Item 70, the fact that he is told that the area of one circle is *four* times the other will predispose the unthinking student to select one of the responses containing the number 4. To a student unfamiliar with the behavior of fractions, the third response in Item 71 will appear to be much less plausible than the fourth. In the next example (72), probably every one of the wrong responses is more enticing to the uninformed than is the correct one.

What we have been saying here can be summed up briefly in the advice: Make the incorrect responses of multiple-choice items as plausible as possible, with the deliberate aim of misleading students of superficial understanding. Such advice might be dangerous in some subjects where test makers, in striving for plausibility in the alternatives, might introduce real ambiguity into the test items. But there is little chance for ambiguity in mathematics. The results of mathematical

operations are either right or wrong. There is little room for differences of opinion, and if the test maker exercises even a minimum of intelligent care, he will run little risk of building ambiguous items.

Suggestion F

Several new-type techniques are not particularly amenable to the measurement of achievement in mathematics but may be used advantageously for instructional purposes.

True-False Tests

The true-false technique has been the object of much criticism in recent years despite its early favor among test makers. Although its general use in achievement testing in mathematics has been somewhat discredited, it is nevertheless a most useful teaching instrument. Much value can accrue from the use of true-false tests if their administration is followed by a general class discussion. This is particularly true in the fields of plane and solid geometry. The following examples are illustrative of items which may be used in this manner.

73. Two right isosceles triangles are congruent if a leg of one equals a leg of the other. ()
74. Two isosceles triangles are similar if any angle of one equals the corresponding angle of the other. ()
75. In any right triangle with unequal legs, the longer leg is shorter than the median to the shorter leg. ()
76. If a segment of a circle contains an inscribed angle of 130° , the center of the circle lies within that segment. ()

The true-false test exists in several variations. The danger in the form illustrated above is that a question chosen may be neither always true nor always false. Because so many statements are neither wholly true nor wholly false, techniques have been devised to provide for a third possibility. In one of these,

each item is preceded by the formula T F S, and the student is instructed to encircle the T if the item is wholly true; the F if it is wholly false; the S if it is sometimes true and sometimes false. In a statement such as "an equilateral quadrilateral is a square," the S should be encircled.

Such a device increases the scope of the test by allowing the use of questions that require closer discrimination, but it is a more difficult test than the plain true-false type.

It should be noted that the rating of a true-false test is usually not best accomplished by counting the correct responses only. The element of chance should be considered, since in such a test it may be a large factor. To score the test more accurately, the number of wrong responses should be subtracted from the number of correct responses, the difference being the score obtained by the pupil.

Similarly, in a test providing three possibilities, one-half the number of wrong responses should be subtracted from the number of correct responses in order to obtain the pupil's score.

In an attempt to increase the value of a true-false test, pupils are sometimes asked to rewrite the false statements correctly. This would seem to have the advantage of counteracting possible impressions made by the wrong statements, as well as of testing the pupil's completeness of understanding of the facts concerned. Unfortunately, however, many false statements that are desirable material for a true-false test might be answered in various ways or would be too difficult for a pupil to correct at the stage of the work when the statements might be most valuable for such use. For example, the false statement, "The sum of the angles of a triangle is four right angles," could readily be corrected to "two right angles," though the pupil might feel that the conclusion should remain the same, and so might choose to change "triangle" to "quad-

rilateral." He might even insert "exterior" before "angles," or make still other changes. Also, the pupil who realized that the statement, "The diagonals of an isosceles trapezoid are perpendicular to each other," was false might not realize that the replacement of "isosceles trapezoid" by "rhombus" would be admissible, or might hesitate to change the conclusion to equality instead of perpendicularity. A teacher can guard against this possibility, to some extent at least, by such a preliminary statement as, "Some of the conclusions drawn from the conditions in the following statements are false. If you find such a statement, mark it 'false,' and rewrite the statement, keeping the same conditions and correcting the conclusion."

Completion Tests

Completion tests also meet with little general use as measures of achievement in mathematics. Some of the more critical considerations involved in their use are as follows:

1. Many of the items usually couched in this form can be more conveniently expressed as short-answer items.
 2. The completion form may be used advantageously when statement of a problem in short-answer form is not convenient.
- Examples:

77. To double the area of a figure, keeping it similar to its original form, the sides must be made _____ times as long.
 78. It is impossible, in general, to find the locus of points equidistant from more than _____ given points.
3. The completion form is a useful technique for originating informal class discussion. A short test of this form may be administered at the beginning of the period and each pupil given time to write his reactions. The ensuing discussion will then be more purposeful and effective. The following examples

are illustrative of test items which may be used in this manner.

79. A quadrilateral is a parallelogram if its diagonals _____.
80. The shortest line-segment from a point to a circle is along a line _____.

Numerous items for such class discussion may also be expressed in the form of direct questions. Examples:

81. What are the two most common ways of proving line-segments equal?
82. What are the principal ways of proving triangles congruent?

4. Completion items may also be used to test pupil information in regard to fundamental definitions. Examples:

83. A(n) _____ is a portion of a circle bounded by two _____ and an arc.
84. A square is a quadrilateral with all its sides _____ and all its angles _____.

5. Too often the answer to a completion item requires nothing more of the pupil than a verbal impression of a textbook definition. Unless he can recall that exact phrasing, the item may be wholly meaningless to him. At other times the item may resolve itself into a test of pupil ingenuity in the use of language alone. Example:

85. _____ may be _____ for _____ in any mathematical expression.

Matching Tests

Matching tests have only infrequently been used in mathematics. The following item illustrates the matching process and also indicates the difficulty which arises in finding mathematical content sufficiently homogeneous to meet the requirements imposed upon this type of technique.

In this exercise the pupil is required to match the items in the

two columns by fitting the correct numbers in the spaces provided.

1. parallelogram	A quadrilateral with two, and only two, sides parallel	(12)
2. corollary	A line drawn from a vertex of a triangle to the midpoint of the opposite side	()
3. bisector	A theorem derived easily from another theorem	()
4. isosceles	A portion of a circle bounded by an arc and its chord	()
5. median	A quadrilateral with all its sides equal	()
6. postulate	A triangle with two equal sides	()
7. rectangle		
8. rhombus		
9. scalene		
10. sector		
11. segment		
12. trapezoid		

Matching tests may be used advantageously as teaching tests where the administration of the tests is followed by class discussion, but their general use for the measurement of achievement in mathematics is somewhat limited.

Miscellaneous Principles of Test Construction

1. Test items designed to measure a pupil's understanding of a *principle* or his ability to *apply* that principle should involve, in general, a minimum of arithmetical computation. Excessive arithmetic computation or manipulation serves only to defeat the purpose of such items. If arithmetical or manipulatory skills are to be tested, these should be tested separately.

2. Some items in the test should involve a necessity for the elimination of irrelevant facts and the combining of relevant facts with the principles essential to the solution of the given problem situation. Actual problems in life situations do not present the essential facts in an orderly fashion, nor are irrelevant elements excluded from the total situation confronting the individual.

3. In general, avoid the use of numerical exercises in which the correct response may be obtained by an incorrect process. This principle may be illustrated by the following test item.

86. Two concentric circles have radii of 3 inches and 5 inches. What is the length of a chord of the larger circle that is tangent to the smaller circle?

The correct response to this item may be obtained simply by adding the numbers 3 and 5. Analysis of test results has shown that pupils often resort to some use of one of the four fundamental operations in obtaining a response to a problem which they have not been able to work correctly. The exercise, "Simplify $x^4 \div x^2$ " provides a similar illustration. The correct response may be obtained by either division or subtraction of the exponents.

4. The difficulty of the items should be specifically adapted to the actual achievement of the pupils to be tested. If the entire group to be tested is considered as classified into a number of levels of achievement, the tests used must contain a due proportion of items to which pupils at each of these levels are just able to respond correctly. Items which will be answered correctly, or incorrectly, by only a negligible proportion of pupils will contribute little to the basic purpose of the examination in discriminating between students at different levels of achievement.

5. Some premium should be placed upon speed as well as upon power in reasoning, by including a sufficient number of items so that most pupils will be given an opportunity to do the maximum amount of work of which they are capable in the time given.

In conclusion, attention should be directed toward the fact that the value of objective techniques for the measurement of achievement in mathematics cannot be ascertained by *a priori* reasoning. Any critical evaluation of these techniques must

be made in the light of the extent to which test items in any form *tend to discriminate* between pupils of varying levels of achievement. Much genuine scientific research must be made in regard to this matter of discrimination if further progress in the field of achievement testing in mathematics is to be attained.

QUESTIONS FOR DISCUSSION

1. Evaluate these statements:
 - (a) "Standardized tests in mathematics are in general limited to the measurement of information and a low level of understanding." Is it probable that some of these tests hold the student responsible for understanding at too high a level?
 - (b) "Standardized tests in mathematics place a distinct premium upon speed of reaction rather than upon power of reasoning."
2. The distinction between "formal" and "functional" testing has been discussed (pp. 346-50). Is it probable that the functional test item measures general intelligence rather than training in a specific mathematical field?
3. Assume that three tests A, B, and C are given to two matched groups of students who have just completed a standard course in elementary algebra. Test A consists of a substantial number of "formal" test items, while test B consists of functional items based on the content of the items in test A. Test C is a comprehensive test of general reading ability. Which of the three correlations between tests A, B, and C would, in your opinion, rank the highest? What educational implication of importance to test makers may be drawn from your answer?
4. How would you propose to obtain the two *matched groups* necessary for the project discussed in exercise 3 above?
5. Evaluate these statements:
 - (a) "Formal" rather than "functional" testing has no place in the instructional program.
 - (b) Objective test items which require only recognition of a correct response have no place in an achievement test in plane geometry.

EXAMINATIONS IN MATHEMATICS

6. One of the important immediate aims of instruction in geometry is the development of ability to work one's way by logical sequence of arguments to the proofs of geometric propositions. To what extent may the traditional, formal proofs of originals be replaced by the numerical short-answer test item in the measurement of achievement in logical demonstration?
7. Two major topics in geometry are "congruence" and "similarity." Examine several standardized tests with a view toward ascertaining their validity in reference to these two major topics.
8. Which of the various testing techniques is most satisfactory for the measurement of the skills, abilities, or understandings included under the general heading of "congruence" and "similarity"?
9. The construction technique illustrated on pages 357-60 is considered useful, not only as a device for testing skill with compasses and straightedge, but as a means for measuring comprehensively all the various aspects of geometry. Under this assumption it might be possible to prepare an achievement test in plane geometry consisting entirely of construction items. Students taking such a test might obtain low scores, not because they do not understand the geometric facts and principles involved in the items, but because of inability to perform the necessary constructions. Such a situation would very likely be found in a course where constructions have not been given due consideration. To what extent would such a situation condition the reliability of the test? the validity of the test? the discriminating power of the individual items?
10. Should each of the six responses in Item 49 (p. 362) be given equal scoring value or should the final response be weighted? Would it be advantageous to give a score-value to only the final response? What correlations would you expect between two sets of scores obtained as follows from a test consisting of a number of items similar to item 49?
 - Score A. Let each response carry a score-value of 1 point.
 - Score B. Let the final response carry a score-value of 6 points, all other responses in each item having no score-value.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

11. Criticize the use of the multiple-response technique for measurement of the specific skills or specific content included in the general topic of "locus" as in Items 63 and 64 (p. 368). Does your criticism apply with equal force to the use of this technique in Item 69 (p. 370)?
12. The general use of the true-false technique has been frequently challenged on the grounds that psychologically it is poor pedagogy to implant misinformation in the mind of the student. Does this consideration necessarily limit the use of true-false questions in mathematics? Does it apply as directly to mathematics as to history, English, or other subjects? Carrying the question beyond the scope of testing: is it pedagogically sound practice while teaching specific operations or content in mathematics to discuss with the pupils the errors that are frequently made, with the intent of teaching them how to avoid such errors?
13. In regard to multiple-response items the authors advise making "the incorrect responses . . . as plausible as possible, with the deliberate aim of misleading students of superficial understanding." In your opinion, to what extent are students actually taught false information by the administration of a test involving such plausible, incorrect responses?
14. Why is homogeneity of content an important consideration in matching exercises? In what respect does the illustrative item on page 376 fail to satisfy this requirement of homogeneity?
15. Discuss: Homogeneity of responses is an absolute requirement in multiple-response items in mathematics.
16. "A test item should be determined essentially by its content and not by its amenability to a prescribed form." This statement is frequently given as an important principle in test construction. Why is it not more generally observed in the construction of standardized tests in mathematics?
17. Would an elementary algebra test consisting of items each involving irrelevant information be essentially a valid test?
18. Examine available standardized tests in Solid Geometry and Trigonometry with particular reference to curricular validity.

CHAPTER VIII

EXAMINATIONS IN ENGLISH

PREVIOUS chapters have emphasized the necessity for a clear definition of objectives before the measurement of progress toward those objectives can profitably be considered, have described the principles and methods which are basic to sound test construction in all fields, and have given specific suggestions for the improvement of examinations of various types. The ways in which most of these suggestions will be helpful to the English teacher should be sufficiently clear without further discussion or specific application. There are a number of problems and techniques of testing, however, which are practically unique to this field of instruction and which have therefore received little or no consideration in the earlier discussions. This chapter will be primarily concerned with these peculiar problems and techniques and is to be considered as distinctly supplementary to the more generalized discussions that have preceded it.

OBJECTIVES IN THE TEACHING OF ENGLISH

The conclusions reached by any one teacher in regard to the ultimate outcomes of instruction in English will depend upon his previous valuations of the aim of education as a whole. For example, the teacher who considers that education is a process of conformity with pressures from without, the absorption of prescribed ideas and methods, will describe the major objectives of the teaching of English in one way; the teacher who maintains that education is a release of the *élan vital*, the burgeoning of an inner nature tending in and of

itself to full development, will describe them in another; the teacher who is convinced that education consists essentially of the reconstruction and reorganization of experience, individual and collective, in a rapidly changing social order, will differ with both.

How directly such differing beliefs and convictions, whether latent or manifest, reflect themselves in the curriculum and in teaching practice may be illustrated by the different types of reading lists which result from the different views of education just described. In the first instance, there is usually a prescribed list of books, arbitrarily selected from the literature of the past; in the second instance, the individual is permitted to select his own reading materials in accordance with the interest or the felt need of the moment; in the third instance, the criterion in the selection is the needs of the individual and the group in the environment of today.

It cannot be too strongly emphasized that any discussion of ultimate objectives, or of examinations in relation to them, is futile until those objectives have been clarified and granted general acceptance. Testing instruments cannot be designed to measure goals of learning that are too nebulous for concise statement or that will be accepted by different teachers only when modified to accord with individual viewpoints.

Fortunately there are certain more immediate outcomes of instruction in English which are fairly common to all these divergent views of ultimate objectives, and which in many instances are susceptible to measurement. This discussion of examinations in English will necessarily be limited to such generally accepted outcomes, as described in the next few paragraphs. While it is not expected that all teachers of English will agree fully with this statement of objectives, any differences between the description here given and that of an individual teacher are likely to be mainly differences in empha-

sis and will be of small consequence as far as the consideration of testing techniques is concerned.¹

The course of instruction in English in schools and colleges has two major divisions: literature and language. A corresponding dichotomy may be made in the objectives of instruction. Those for literature may be termed the assimilative or literal, since they concern the exploration of the written expression of other men and women. Those for language may be termed the reproductive or creative, since they are the goals of the learner for the expression of his own thoughts and emotions.

The first category includes: attainment of ability to read literary materials with facility and understanding; development of critical judgment and appreciation of literature; enlarged acquaintance with literature and with literary history; broadening of experience vicariously through reading; formation of desirable attitudes toward reading, namely, increased appetite and taste for what is good in literature; and attainment of some competency in the use of the resources of libraries.

The second category includes: attainment of the ability to speak and write intelligibly, agreeably, and effectively, and development of desirable attitudes toward the translation of experience into spoken and written words, namely, the desire to speak and write for the satisfaction of creator or of audience, or of both. The first of these objectives involves, among other things, familiarity with the conventional tools of speech and writing as studied under the usual classifications of grammar, punctuation, spelling, etc.; the formation of correct habits in the use of these tools; and, at a higher level, the development of ability to organize the results of thought and experience into

¹ See p. 385.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

larger units of effective expression. This latter aspect differs from the preceding two in the same manner that speaking and writing characterized by power differ from speaking and writing characterized by mere correctness, and involves the coherent presentation of one's own ideas.

These objectives, however they may be emphasized, are main goals for the learning of English in schools because they are, except in so far as the term "English" is used to include the scientific researches of scholars, the goals of learning English throughout life. But it is clear that some of the objectives stated are dependent and conditional upon the attainment of others. For example, desirable attitudes toward literature can, in general, be attained only through skill in silent reading, and most genuine and abiding knowledge of literature springs from the establishment of an appetite and taste for it. While no sharp lines can be drawn between the learning processes involved in their attainment, the objectives described may be conveniently tabulated thus:

<i>A</i>	<i>B</i>
<i>Assimilative or Literal</i>	<i>Reproductive or Creative</i>
1. Reading abilities	1. Speaking and writing abilities
2. Critical abilities	
3. Acquaintance with literature and literary history	a. Skill in correct language usage
4. Vicarious experience	b. Power of expression
5. Reading attitudes	2. Speaking and writing attitudes
6. Ability to use the library	

HOW TO EVALUATE ACHIEVEMENT

In the remaining discussion an attempt will be made to determine what means of evaluation will give teacher or learner, or both, the fullest information about the progress of the

students, individually and collectively, toward these several goals.

The Assimilative or Literal Objectives

The problem of measuring the results of instruction is perhaps farther from satisfactory solution in literature than in any other field, particularly at the high-school level. This fact results from certain peculiarities of this field of instruction which place definite restrictions upon the examiner. The subjectivity of the content is one impediment to the development of adequate testing techniques. The wide variability in materials and methods of instruction is another. The lack of agreement upon objectives, already pointed out, is a third and very serious obstacle. Furthermore, measurement of the extent to which any objective has been realized depends upon the examiner's ability to identify the specific elements of behavior in the student which are indicative of the attainment of that objective. Unfortunately, neither test authorities nor teachers of English are yet prepared to describe all of the desirable objectives in terms of specific and readily observable student behavior. As a result, relatively little progress has been made in the development of techniques for their measurement.

For these reasons, the following discussion must be devoted principally to the testing of those outcomes which are fairly concrete and for which practicable and reliable measuring devices have been developed. Only a few suggestions can be offered concerning the less tangible objectives in the preceding tabulation, and those suggestions will necessarily be comparatively vague and arbitrary. However, the impracticability of attempting to measure *all* the results of instruction should not interfere with attempts at better evaluation of the outcomes that are measurable.

1. Reading abilities**2. Critical abilities**

Because of the overlap in these two types of abilities, their measurement can best be considered simultaneously in order to avoid repetition.

The measurement of the pupil's general reading ability, particularly with reference to informational materials in the content subjects, is not the special concern of the English teacher, nor need he attempt to prepare his own examinations for that purpose. Many standardized tests are available for the measurement of this general ability,² as well as much published information in regard to what it involves.³ Some of these tests are specially designed to yield, not only measures of comprehension in silent reading, but also evidence of achievement in several different phases of reading ability of the work-study type.⁴ Since the attainment of all objectives of the assimilative or literal type is dependent upon the establishment of the abilities measured by these tests, the teacher of English is advised to become thoroughly familiar with them and to use them in discovering the pupil's level of habitual skill in reading. For the measurement of those reading objectives that are unique to the study of literature, however, the teacher will often have to prepare his own examinations. The subsequent discussion will be concerned only with the construction of tests for that purpose.

² P. V. Sangren, "Improvement of Reading through the Use of Tests." Bulletin of Western State Teachers College, 27, No. 1, Kalamazoo, Michigan: Western State Teachers College, 1931.

³ See the selective bibliography of material which has appeared in the last four years prepared by the Educational Records Bureau, 437 West 59th Street, New York City, Jan. 15, 1935.

⁴ Thus, the Iowa Silent Reading Test gives separate scores for paragraph meaning, word meaning, selection of central idea, sentence meaning, location of information, rate, and total comprehension. Such tests are more useful to teacher and pupil for analysis of difficulties, and more helpful for remedial instruction, than the tests which yield only a score on speed and comprehension.

*Special Problems in the Reading and Appreciation of
Literature*

As has just been implied, the reading of literature requires of the pupil certain insights and abilities that are not involved in reading informational materials of the content subjects. For example, requisite to the comprehension of certain literary materials, particularly those of past periods, is familiarity with conventional modes of literary expression. To far greater extent than in ordinary reading, the pupil must be able to comprehend such devices as figures of speech, symbolism, personification, puns, comparison and contrast, historical, classical, and mythological allusions and analogies, allegory, etc. In poetry, where inversions and irregular grammatical structures frequently occur, knowledge of verse patterns, types of meter, etc., may assist the pupil in following the thought of the stanzas. These, and many other literary methods and devices which will suggest themselves to the teacher, should be given special attention in tests of reading comprehension of literature. They should not, of course, be overemphasized to the exclusion of subtler considerations. The pupil's perception of the mood of a selection, for instance, deeply affects both his understanding of and his emotional reaction to the passage. Although his emotional response cannot be measured, it is possible to test the pupil's recognition of the mood or feeling expressed in a selection and of the means by which the writer gained that effect. The pupil should be able to discern whether the tone of a passage is, say, facetious or serious, animated or matter-of-fact; he should be able to distinguish satire, humor, and burlesque; in cases where the writer's own attitude is particularly evident, the pupil should be able to detect its tenor. All of these factors should be considered in the construction of test questions.

At a higher level, if the pupil is to gain the greatest profit and enjoyment from his reading, he should be able to consider a literary composition in its totality. He should be aware, not only of sentence meanings, but of the general import, direction, or purpose of the poem, play, essay, or novel read — to sense, in so far as he is able, the author's intention. A very high degree of understanding of this type cannot be expected of the average high-school pupil, but its furtherance should be an aim of instruction. Although the testing of this phase of literary comprehension is complicated by the fact that the passages included in the examinations either must be very brief or else must be excerpts from a longer work, some questions of this type may be asked where the nature of the passage permits. Shorter poems lend themselves particularly well to testing on this point.

In developing his critical judgment of literature the pupil should learn to recognize through analysis, comparison, and appraisal in what particular a given passage is superior or inferior, and should become critically aware of specific merits and demerits of form and style. Those elements of form and expression which the high-school pupil may be expected to appraise most adequately are, fortunately, also those which can be tested most reliably. Examples of such points are: individuality or commonness of diction, clarity or confusion of imagery, simplicity or complexity of style, coarseness or subtlety of humor, strength or weakness of character portrayal, smoothness or irregularity of rhythm, etc. The high-school pupil may also be expected to detect such deeper qualities as sentimentality or insincerity on the part of the author, and to a certain extent to judge the fairness or unfairness of the writer's own attitude toward the characters or ideas presented.

*The Essay Type of Test as Applied in the Measurement
of Comprehension and Appreciation of Literature*

One method of measuring the outcomes which have just been identified is to present to the pupil a number of literary selections which are likely to be unfamiliar to him, and then, in relation to these selections, to ask him questions of the kinds which have been suggested in the preceding paragraphs. The pupil might be presented, for example, with the following selection.

- (1) 'Twas on a May-day of the far old year
- (2) Seventeen hundred eighty, that there fell
- (3) Over the bloom and sweet life of the Spring,
- (4) Over the fresh earth and the heaven of noon,)
- (5) A horror of great darkness, like the night
- (6) In day of which the Norland sagas tell, —
- (7) The Twilight of the Gods. The low-hung sky
- (8) Was black with ominous clouds, save where its rim
- (9) Was fringed with a dull glow, like that which climbs
- (10) The crater's sides from the red hell below.

Some essay-type questions which might be asked to discover whether or not the pupil fully comprehends this selection are:

1. Of what kind of event is the poet speaking?
2. What is the meaning of "ominous" as used in line 8?
3. What are the "Norland sagas" (line 6)?
4. What was the source of the "dull glow" (line 9) upon the sky?
5. To what do the last two lines refer?

The following selection and questions further illustrate this kind of testing.

- (1) She glanced through the fly-specked windows of the
- (2) most pretentious building in sight, the one place which
- (3) welcomed strangers and determined their opinion of the
- (4) charm and luxury of Gopher Prairie — the Minniemashie

EXAMINATIONS IN MAJOR SUBJECT FIELDS

- (5) House. It was a tall lean shabby structure, three stories
(6) of yellow-streaked wood, the corners covered with
(7) sanded pine slabs purporting to symbolize stone. In the
(8) hotel office she could see a stretch of bare unclean floor,
(9) a line of rickety chairs with brass cuspidors between,
(10) a writing-desk with advertisements in mother-of-pearl
(11) letters upon the glass-covered back. The dining-room be-
(12) yond was a jungle of stained table-cloths and catsup
(13) bottles.
(14) She looked no more at the Minniemashie House.
6. What would you say is the writer's own opinion of the Minniemashie House?
 7. What method does the writer employ predominantly in attempting to achieve the desired effect?
 8. Is that effect successful, and upon what do you base your judgment?
 9. What does the writer imply in his reference to "the charm and luxury of Gopher Prairie"? What literary technique is exemplified in this particular part of the passage; that is, what name do we give to this method of conveying one's meaning?
 10. What is the meaning of the word "rickety" in line 9?
 11. What is the meaning of the word "symbolize" in line 7?
 12. The last line is an example of what formal means of achieving emphasis?

This method is superior to the "discuss this passage" procedure in that it identifies for all pupils the points on which they are expected to respond. Such identification assures that no pupil will neglect, through sheer oversight, some element about which the teacher wishes to measure his understanding; it also reduces the difficulty of scoring by permitting the teacher to assign definite values to specific elements, and by insuring that the pupil's responses to those elements will be found sufficiently isolated to permit ready and adequate grading of each one. Regardless of the procedure used, however, the scoring of most essay questions in literature is certain to be highly sub-

jective. The suggestions for scoring essay tests presented in Chapter V on the natural sciences should prove helpful to the teacher of literature also. These suggested procedures include the preparation of a set of specifications in which the nature of the desired response is definitely determined for each question and in which a scale of possible scores that will permit evaluating all responses, from very good to very poor, is established for each item. The purpose of the test as a whole and of each item individually must be kept well in mind during the scoring. If the examination is intended to measure the pupil's ability in reading and understanding literature, he should not be penalized for faults of composition, such as misspellings, errors in grammar, etc. Although such errors may be checked or corrected on his paper, they should detract as little as possible from his score in reading comprehension.

In framing such essay questions, care must be taken to indicate clearly the point under consideration in each item. For example, if one wished to test for the pupils' realization of the general untidiness of the Minniemashie House, he should not ask such a question as, "What kind of place was the Minniemashie House?" Some pupils would probably answer that it was a hotel; others, that it was a "tall lean shabby structure"; still others, that it was poorly furnished. The expected answer would not be likely to appear unless a pupil repeated the entire description given in the passage. Likewise, in testing on formal devices one should be sure to indicate clearly the element under consideration and also the manner in which it is to be identified. For instance, in a passage of poetry one should not ask merely, "What literary device does the poet employ?" if the reference is to, say, the poet's use of feminine rhyme; the pupil will simply be confused in trying to decide to what element of poetic structure the question refers. Or in the illustrative passage quoted above, it would be insufficient to ask, "Of

EXAMINATIONS IN MAJOR SUBJECT FIELDS

what is the last sentence an example?" if the expected answer were "Understatement."

Further cautions in the choice of selections and of points to be tested are given in the following discussion of the objective variant of this form of test exercise. These cautions apply with equal force to the subjective or essay variety, and should be closely observed.

An Objective Type of Test for Measuring Comprehension and Appreciation

A number of forms of tests have been devised for the objective measurement of those aspects of silent reading comprehension that are unique to the study of literature. One of the most promising of these techniques, which is an adaptation of the type of exercise just discussed in reference to essay examinations, may be illustrated by the following examples taken from materials prepared for the Cooperative Literary Comprehension Test.

- (1) Dear Harp of my Country! in darkness I found thee,
 - (2) The cold chain of silence had hung o'er thee long,
 - (3) When proudly, my own Island Harp! I unbound thee,
 - (4) And gave all thy chords to light, freedom, and song!
 - (5) The warm lay of love and the light note of gladness
 - (6) Have wakened thy fondest, thy liveliest thrill;
 - (7) But, so oft hast thou echoed the deep sigh of sadness,
 - (8) That even in thy mirth it will steal from thee still.
13. This passage seems to refer to 1 England, 2 France,
3 Germany, 4 Ireland, 5 Wales..... ()
 14. The writer is discussing 1 his falling in love, 2 the lib-
eration of his country, 3 the liberation of the press,
4 a religious revival, 5 a revival of poetry..... ()
 15. In this passage the writer employs 1 a unique meter,
2 climax, 3 feminine rhymes, 4 internal rhymes, 5 rep-
etition..... ()

EXAMINATIONS IN ENGLISH

16. The last two lines indicate that mirth 1 conquers all, 2 has returned, 3 is pointless, 4 must be accompanied by sorrow, 5 steals away our hearts. ()
17. "Cold chain of silence," in line 2, refers to 1 imprisonment, 2 lack of poetry, 3 loss of liberty, 4 loss of religious freedom, 5 unconfessed sins. ()
 - (1) Old Behrman was a painter who lived on the ground
 - (2) floor beneath them. He was past sixty and had a Mi-
 - (3) chelangelo's Moses beard curling down from the head
 - (4) of a satyr along the body of an imp. Behrman was a
 - (5) failure in art. Forty years he had wielded the brush
 - (6) without getting near enough to touch the hem of his
 - (7) mistress's robe. He had been always about to paint a
 - (8) masterpiece but had never yet begun it. For several
 - (9) years he had painted nothing except now and then a daub
 - (10) in the line of commerce or advertizing. He earned a little
 - (11) by serving as a model to those young artists in the colony
 - (12) who could not pay the price of a professional. He drank
 - (13) gin to excess, and still talked about his coming master-
 - (14) piece.
18. Behrman seems to have been 1 boastful, 2 brutally frank, 3 taciturn, 4 downcast, 5 reserved. ()
19. In appearance he was 1 handsome, 2 strikingly robust, 3 queer, 4 spiritual, 5 commonplace. ()
20. His work was 1 excellent, 2 generally good, 3 occasionally good, 4 marked by genius, 5 poor. ()
21. "To touch the hem of his mistress's robe," line 6, means to 1 become an artist, 2 gain wealth, 3 make friends with women, 4 marry the woman he loved, 5 win fame. ()
22. The reference to a satyr conveys an idea of 1 friendliness, 2 sullenness, 3 grossness, 4 beauty, 5 spirituality. ()
23. The writer's own attitude toward Behrman is 1 approving, 2 pitying, 3 highly disapproving, 4 bitterly contemptuous, 5 matter-of-fact. ()
 - (1) I admitted of course that Virgil in spite of his genius had
 - (2) a hardness and an old glitter which resembled rather the

- (3) brilliance of a cut diamond than the soft grace of a flower.
 (4) Certainly I admitted this; the mere admission of it would
 (5) knock the breath out of anyone who was arguing.
 (6) From such talks my friends went away sad. The con-
 (7) clusion was too cruel. It had all the cold logic of a
 (8) syllogism (like that almost brutal form of argument so
 (9) much admired in the Paraphernalia of Socrates). For if:—
 (10) Virgil and Homer and Pindar had all this grace, and
 (11) pith and these sallies, —
 (12) And if I read Virgil and Homer and Pindar,
 (13) And if they only read Mrs. Wharton and Mrs. Humphry
 (14) Ward,
 (15) Then where were they?
 (16) So continued lying brought its own reward in sense of
 (17) superiority and I lied more.
24. The writer's admiration for Virgil is 1 fanatical,
 2 genuine, 3 half-hearted, 4 misunderstood, 5 pre-
 tended. ()
25. The use of the word "Paraphernalia" in line 9 gives a
 touch of 1 burlesque, 2 poetry, 3 sentimentality,
 4 stupidity, 5 vulgarity. ()
26. A "syllogism" (line 8) is 1 an accurate description,
 2 a clear definition, 3 a form of literature, 4 a form of
 reasoning, 5 a type of mind. ()
27. "Sallies" as used in line 11 means passages of 1 com-
 plimentary language, 2 figurative language, 3 humor,
 4 personal feeling, 5 sound thoughts. ()
28. The writer is trying to be 1 earnest, 2 impartial,
 3 poetic, 4 scholarly, 5 witty. ()
29. He admitted the hardness of Virgil 1 bitterly, 2 for
 the sake of argument, 3 voluntarily, 4 grudgingly,
 5 without special interest. ()
- (1) The Lord my pasture shall prepare,
 (2) And feed me with a shepherd's care;
 (3) His presence shall my wants supply,
 (4) And guard me with a watchful eye;
 (5) My noonday walks he shall attend,
 (6) And all my midnight hours defend.

EXAMINATIONS IN ENGLISH

- (7) When in the sultry glebe I faint,
 (8) Or on the thirsty mountain pant;
 (9) To fertile vales and dewy meads
 (10) My weary wandering steps he leads:
 (11) Where peaceful rivers, soft and slow,
 (12) Amid the verdant landscape flow.
30. The passage suggests the 1 doxology, 2 *Gloria in Excelsis*, 3 parable, 4 nineteenth psalm, 5 twenty-third psalm. ()
31. "Glebe" in line 7 means 1 a desert, 2 a field, 3 a forest, 4 an oasis, 5 a road. ()
32. "Verdant" in line 12 means 1 ample, 2 green, 3 luxurious, 4 wealthy, 5 wide. ()
33. This passage expresses 1 brotherhood, 2 faith, 3 patience, 4 penitence, 5 a wish to serve. ()
34. The imagery is 1 commonplace, 2 grotesque, 3 startling, 4 unique, 5 vulgar. ()

In constructing such items, the teacher must guard particularly against ambiguity. As in the essay questions previously discussed, the purpose of each item, i.e., the element for which it is testing, should be perfectly clear to the pupil. The items should be stated so concisely and completely that the pupil will not become confused by having to ponder upon what the test constructor had in mind. A very simple illustration of grammatical ambiguity would be an item in which the unidentified pronoun "he" was used with reference to a passage that concerned more than one male character. As another example, an ambiguously phrased item on "attitude" might be answered by one pupil with reference to the author of the selection and by another pupil with reference to a character in the selection. Also, names not mentioned in the passage should not be used in the items, and any words or phrases from the passage that are quoted in the item should be quoted exactly to avoid misinterpretation. Such "slips" can readily be avoided, but are easily overlooked by the test constructor himself, who has the

intended meaning so clearly in mind. Much more complex instances of ambiguity could be cited. Consider the following passage and items.

Schiller! that hour I would have wished to die,
If through the shuddering midnight I had sent
From the dark dungeon of the tower time-rent
That fearful voice, a famished father's cry —
Lest in some after moment aught more mean
Might stamp me mortal! A triumphant shout
Black Horror screamed, and all her goblin rout
Diminished shrunk from the more withering scene!

35. The poet would wish to die because the effect was
1 false to truth, 2 insane, 3 successful, 4 weakening,
5 wicked..... ()
36. The withering effect was on the person's 1 character,
2 feelings, 3 hopes, 4 religious beliefs, 5 wicked
lusts..... ()

The pupil could not respond to Item 35 without first inquiring, "What effect?" or "Effect upon whom?" In Item 36 likewise, the meaning and reference of the words "withering effect" and "the person's" are so doubtful that the item is unanswerable without further clarification. Similar indeterminate or inaccurate references to the content of a passage may render an item entirely non-functioning. Precise wording is equally important in the *responses* of a multiple-choice item. Careless phraseology may make an intended wrong response appear to coincide with the passage, or may alter the meaning of the correct response enough so that the more astute pupils, particularly, will not be satisfied to accept it as a valid interpretation of the passage. All of the wrong responses should be carefully checked to insure that neither they nor the passage itself could be interpreted in a manner that would make any one of them an acceptable answer.

The latter admonition is especially important where the

selection quoted is an excerpt from a longer composition. Often the interpretation of an excerpt will differ from the interpretation of the same passage in its original context. The test constructor must take cognizance of this fact. He should ask no questions which cannot be answered on the basis of the material actually before the pupil, and he should phrase the items in such a way that the pupil will not be misled into making interpretations which could be justified by the excerpt but which would be considered definitely incorrect on the basis of the original context.

In vocabulary items, there is danger, unless particular care is taken in their construction, that the "incorrect" responses may be justifiably construed by the pupil as correct according to his own subjective interpretation of the passage. It is relevant to mention also that, because of this subjective factor, the responses to vocabulary items should not demand too fine a discrimination in word meanings; furthermore, the words constituting the responses should be simple enough so that they themselves will not present vocabulary difficulties to the pupil.

The matter of interpretation of the selection is, of course, of fundamental importance. Since the objective method of scoring does not allow for the variations in personal opinion that frequently arise among different readers of a literary passage, the test items themselves must present interpretations that are likely to be considered valid by most readers. The surest safeguard against the inclusion of controversial or indeterminate items is to have one or more other persons review the test critically, comparing their own interpretations with those expressed in the items, and to discard or revise all items on which disagreements cannot be readily resolved. One should avoid testing on highly abstruse literature in which interpretations are sure to be influenced by subjective factors. Items suffi-

ciently difficult for measuring the abilities of the superior pupils can be constructed without recourse to such materials.

One should be careful, also, not to ask for implications beyond the discernment that may be expected of pupils at the level tested. However, the other extreme of stressing obvious and purely literal meanings is equally undesirable. Some easy items of the latter type will have to be included, of course, in order to measure the achievement of the inferior pupils, but even in such items the test constructor need not and should not test for mere trivialities or for very minor details that have little bearing upon the central point of the passage. For example, in the first passage quoted on page 393, it would be extremely easy but futile to construct an item on the fact that Behrman had done some commercial art work, as it would test for the fact that some young artists in the colony apparently were very poor. One should try to recognize the reading problems of the pupils and frame items with those difficulties in mind. An attempt should be made to measure as many as possible of the behavior-evidences related to the objectives concerned, i.e., ability to recognize and to evaluate specific points of the passage which exemplify the various elements enumerated on pages 387-88.

Preliminary tryout of tests of this kind is very desirable, because it is often difficult to anticipate correctly the thought-processes of the pupils in reading a specific passage and to avoid all ambiguities in the items. An item which seems quite obvious to the teacher, in the light of his superior knowledge, may appear extremely difficult or even unanswerable to the pupil who knows nothing of the origin and history of the selection. Extensive tryout is, of course, impracticable for the typical classroom teacher, but if carefully constructed test exercises are used repeatedly with different groups, they can be revised and improved on the basis of previous administrations.

Selection of Test Materials

The choice of the selections employed in the test is quite as consequential as the construction of the items. It is highly important that the passages used be unfamiliar to the pupil, particularly that they should not have previously been discussed in class; otherwise the pupil may answer simply upon the basis of his recall of what the teacher or commentators have said and not upon his own understanding of the selection. The construction of a test of this kind requires, of course, a very thorough familiarity on the part of the teacher with a very large number of literary selections, since it is difficult to find short passages that are sufficiently rich in reading problems to adapt themselves well to this type of testing. The use of long passages that contain only a few salient points for testing will merely waste the pupil's time in superfluous reading. An excerpt from a long composition should present a fairly complete unit of content that is meaningful without reference to the original context and that does not have a too "decapitated" appearance. Ambiguous fragments at the beginning or the end of an excerpt may lead the pupil astray in his interpretation of it.

If the examination is intended to measure general achievement in comprehension or judgment, or both, the selections included should represent as wide a variety of types of literary materials as the pupil is likely to encounter in general reading. Poetry and prose should be well balanced; the broad categories of composition — narration, description, exposition — should be represented, and the selections should offer a range of style, mood and treatment. If the test is intended to measure only certain phases of the abilities involved in these major objectives, the type of selections used may be varied according to the emphasis desired.

Instructional Values of Test Materials

This type of examination has distinct teaching values. It is possible, by careful item construction, to bring to the pupil's attention important points that he might otherwise have overlooked in the passage, and to force him to consider their meaning and implications. By thus helping him to realize the significances underlying a piece of literature, this testing device may increase his appreciation and enjoyment of that particular passage, may sharpen his awareness of similar possibilities in other compositions, and may help to stimulate his interest in further exploration of the field of literature. In this regard, it is relevant to caution the teacher once more against ambiguity, triviality, and doubtful interpretations in item construction. Such structural defects in the items employed will merely exasperate the pupil and dull his interest.

3. Acquaintance with Literature and Literary History*Testing for Acquaintance with Literature*

In the type of objective examination just discussed, questions can be included which will test familiarity with particular authors through allusion to the characteristics of their writings. For example:

37. This passage suggests the writings of 1 Browning,
2 Milton, 3 Pope. ()
38. The attitude toward the common people expressed in
this passage is that of 1 Johnson, 2 Pope, 3 Words-
worth. ()

In order to answer this type of item the pupil must have a first-hand familiarity with the writings of the authors cited, a familiarity that can be gained only through alert and thoughtful reading. For that reason, this type of item is superior to

the following, in which a general question is asked without reference to any particular passage.

39. His obscure style has frequently aroused comment:
 1 Addison, 2 Browning, 3 Dickens, 4 Dryden,
 5 Swift..... ()

The weakness of this latter type of item is that, because the range of information adaptable to it is limited, it tends to test for the mere *clichés* of literary scholarship. In item 39, for example, the adjective "obscure" has been glibly applied to the poetry of Browning for so long and in such innumerable classroom discussions and textbooks in literature, that the pupil is likely to link it automatically with the name "Browning" even though he has never read a page from that poet. This form of item may be useful in testing for opinions that the pupil may reasonably be expected to have formed for himself without teacher or textbook repetition, but if information upon which the pupils have previously been drilled is introduced into the test, it should be disguised by new phraseology so that the rote-learner will not be able merely to parrot the correct answer.

In extensive testing for *range* of acquaintance with literature, items of the type of examples 37 and 38 are not very practicable, since they require the inclusion of literary selections in the test. A better test for general usage is the straight multiple-choice type of examination. It should be emphasized that the real objective in this phase of instruction is to encourage the pupil to become *directly* acquainted with authors and their works rather than merely to be able to associate the name of a literary work with the name of its author or to be able to repeat statements about the work that have been made by other people. In testing for range of literary acquaintance, therefore, "who wrote what" types of questions, which test for little more than verbal learning, should be used sparingly if at all.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

Rather, the questions should be of such a character that the pupil is likely to respond correctly only if he has actually read and understood the piece of literature involved. This desirable characteristic can be secured by basing the questions upon quotations; reference to setting, plot, characters, etc.; reference to general significance in literary history, to some historical or social aspect of the author's writings; and so on. Examples of various approaches are shown in the following items, taken from the Cooperative Literary Acquaintance Test, Form 1935.

40. The most significant characteristic of Lady Macbeth is 1 fear of moral torment, 2 patriotism, 3 ambition for her husband, 4 gentleness of nature, 5 generosity toward her enemies. ()
41. "With a long grey beard and glittering eye," describes 1 Salome, 2 Cyrano de Bergerac, 3 Tristram Shandy, 4 the Ancient Mariner, 5 Ichabod Crane. ()
42. "But whosoever shall smite thee on thy right cheek, turn to him the other also," is quoted from 1 Abraham, 2 David, 3 Paul, 4 Jesus, 5 Isaiah. ()
43. The play that discusses with convincing fairness the eternal clash between capital and labor is 1 *Cyrano de Bergerac*, 2 Galsworthy's *Strife*, 3 *Riders to the Sea*, 4 *The Blue Bird*, 5 *Arms and the Man*. ()
44. Synonymous with loutish, prankish youth is the name 1 Tony Lumpkin, 2 Malvolio, 3 Fabian, 4 Mrs. Malaprop, 5 Sir Toby Belch. ()
45. Howells' *Literary Friends and Acquaintances* is a classic account of the silver age of 1 Philadelphia, 2 San Francisco, 3 New Orleans, 4 New York, 5 Boston and Cambridge. ()
46. Of the following writers, the most radical American experimenter in poetic form and substance was 1 Aldington, 2 Pirandello, 3 Vachel Lindsay, 4 Rupert Brooke, 5 Susan Glaspell. ()
47. The poet who may best be compared with Poe is 1 Arnold, 2 Santayana, 3 Masefield, 4 Jeffers, 5 Coleridge. ()

EXAMINATIONS IN ENGLISH

48. In *The Divine Comedy*, Beatrice represents 1 retribution, 2 comedy, 3 Purgatory, 4 the seven deadly sins, 5 heavenly love..... ()
49. Don Quixote mistook for mighty armies 1 a grove of trees, 2 a group of merchants, 3 windmills, 4 two flocks of sheep, 5 a rocky field..... ()
50. The psychology of love and marriage is the chief concern of 1 Barrie, 2 Rostand, 3 Schnitzler, 4 Shaw, 5 Galsworthy..... ()

If, as in the case of the Cooperative test, the examination is intended to measure general acquaintance with the whole field of literature and to be administered to varied groups, many of the items included necessarily will not be very searching. Some must be very simple and require little discrimination between writers and works, in order to measure the achievement of pupils who have had little or no instruction in literature and have done a very limited amount of "free" reading. For example:

51. Which Shakespearean play is among the greatest tragedies on the theme of romantic love? 1 *Hamlet*, 2 *Troilus and Cressida*, 3 *Romeo and Juliet*, 4 *The Merchant of Venice*, 5 *Twelfth Night*..... ()
52. Hawthorne's novel which deals with the expiation of a sin is 1 *Twice-Told Tales*, 2 *The Scarlet Letter*, 3 *The Alhambra*, 4 *Omoo*, 5 *The Merchant of Venice*..... ()

These two items test only for very superficial information. The items intended to measure the achievement of the superior pupils, however, should attempt to test for first-hand and more thorough knowledge rather than for mere verbalism and hearsay. Examples 40-50 above represent various degrees of difficulty of this type of item.

Some pupils will arrive at the correct response to many items by a process of elimination. This is not necessarily a defect of the test. The elimination method of arriving at the correct

response is quite justifiable in a general test of this nature, since the ability to detect the wrong responses indicates acquaintance with the portions or elements of literature which they represent. However, the wrong responses should be alternatives that will appear plausible to the uninformed pupil, and no irrelevant clues to the correct response should be included. (See discussion on pp. 67-72.) Consider the following item.

53. "He cut a rope from a broken spar,
And bound her to the mast," is quoted from 1 *Serenade*, 2 *The Wreck of the Hesperus*, 3 *The Raven*,
4 *Evangeline*, 5 *Kubla Khan*. ()

The testee can respond to this item simply on the basis of the titles given. The word "Wreck" in the correct response will naturally be linked with the words "broken spar" in the quotation, particularly since the name "the Hesperus" sounds appropriate to a ship and since "mast" is also mentioned in the quotation. This clue is further strengthened by the fact that there is nothing in any one of the other four titles that even remotely suggests ships or shipwreck. Thus the pupil can readily select the right answer to this item without ever having seen or heard of any of these five literary selections. In the next example, the pupil needs merely to sense that "Nibelungenlied" has the general characteristics of a German word.

54. The *Nibelungenlied* is the national epic of 1 Persia,
2 Spain, 3 Greece, 4 India, 5 Germany. ()

The defect of non-functioning content (see pp. 73-81) is another common weakness of this type of item, and frequently results in confusion concerning what particular bit of information is actually being measured. For example:

55. The triumph of a woman who found in grinding toil a romantic adventure is delineated in Edna Ferber's
1 *Seventeen*, 2 *The Little French Girl*, 3 *So Big*,
4 *Twilight Sleep*, 5 *Cimarron*. ()

This item apparently purports to measure the pupil's knowledge of the content of *So Big*, and if he actually has read the book, the item will probably function as intended. On the other hand, even though he has never read any of these books, if he merely happens to know that of the five listed only *So Big* and *Cimarron* were written by Edna Ferber and if he has seen the moving picture version of the latter, he will certainly select response 3 as the correct answer. When answered by this thought-process, the item tests merely the knowledge that Edna Ferber wrote *So Big*. It would be better, therefore, to clarify the wording of this item so that it will really test for a predetermined phase of information, either the superficial knowledge of "who wrote what" or the deeper acquaintance with the content of the book, in order to know the significance of the measure obtained from it. The following suggests a possible phraseology for each of the two approaches:

56. Edna Ferber wrote 1 *Seventeen*, 2 *The Little French Girl*, 3 *So Big*, 4 *Twilight Sleep*, 5 *Miss Lulu Bett*. . . . ()
57. The triumph of a woman who found in grinding toil a romantic adventure is delineated in Edna Ferber's 1 *Cimarron*, 2 *So Big*, 3 *Buttered Side Down*, 4 *The Girls*, 5 *Cheerful — by Request*. ()

Item 52 above exhibits a similar weakness.

Testing for Knowledge of Literary History

The multiple-choice type of item can be used in testing for knowledge of literary history as distinct from literature itself, such as the biographies of writers, the dates and characteristics of important periods of literary production, chronological relationships between writers and between the works of one or of several writers, and so on. In this phase of measurement, again, trivialities should not be given undeserved attention; for instance, in testing on biography, one should attempt to

EXAMINATIONS IN MAJOR SUBJECT FIELDS

select circumstances which seem to have had some bearing upon the author's work or which are otherwise of particular importance and interest. The following illustrations from various Iowa Every-Pupil Tests will merely suggest the variety of questions that may be asked. In a teacher-made examination covering an assigned list of writers recently studied by the group to be tested, more specific biographical details may be included, but these should be expressed in a manner different from the phraseology of the textbook and of class discussion. (See pages 81-96.) Each question should be checked carefully for ambiguities that would make its application uncertain.

58. His interest in Italian freedom appears in his writings:
 1 Browning, 2 Chaucer, 3 Milton, 4 Shakespeare,
 5 Thackeray. ()
59. His ill health and partial deformity affected his whole
 outlook on life: 1 Shakespeare, 2 Dryden, 3 Pope,
 4 Dickens, 5 Spencer. ()
60. He was a rebel against established society: 1 Addison,
 2 Chaucer, 3 Johnson, 4 Shelley, 5 Shakespeare. . . . ()
61. His poems were attacked by critics in a way bitterly
 resented by his friends: 1 Addison, 2 Scott, 3 Keats,
 4 Chaucer, 5 Milton. ()
62. He wrote an important long poem describing his
 travels: 1 Burns, 2 Byron, 3 Tennyson, 4 Milton,
 5 Dryden. ()
63. He came of a poor farming family: 1 Cooper, 2 How-
 ells, 3 Poe, 4 Whitman, 5 Whittier. ()
64. He was a good deal of a vagabond: 1 Emerson,
 2 Holmes, 3 Longfellow, 4 Lowell, 5 Whitman. . . . ()

In order to discourage mere rote learning of dates, chronology may be tested by requiring the pupil to relate one author to another, or to "date" writers and compositions by reference to important historical events or periods. This method has the merit of bringing more forcibly to the pupil's attention the social significance of literature, the influence of contemporary

conditions upon a writer's production, and the differences or similarities in the work of writers of the same period. Illustrative items from the Iowa Every-Pupil Tests are cited below.

65. He began writing after the Civil War: 1 Harte,
2 Bryant, 3 Hawthorne, 4 Irving, 5 Whitman. ()
66. Of the following, he was the first to write: 1 Cooper,
2 Hawthorne, 3 Holmes, 4 Poe, 5 Lowell. ()
67. He wrote at the same time as Shelley: 1 Swift, 2 Addison,
3 Keats, 4 Dryden, 5 Tennyson. ()
68. He wrote during the reign of Queen Anne: 1 Brown-
ing, 2 Burns, 3 Shelley, 4 Swift, 5 Thackeray. ()
69. Of the following writers, the most recent was:
1 Burns, 2 Byron, 3 Eliot, 4 Spenser, 5 Scott. ()

Again, if the test covers a limited list of writers studied in the course of instruction, the questions may deal with more detailed and specific points.

In the construction of tests of acquaintance with literature and literary history, particularly, there are a few possible variations in mechanical form which will suggest themselves to the teacher. Of these the most satisfactory is the matching exercise; it can sometimes be used with greater economy than the multiple-choice type of item. Since, however, the two forms have fundamentally the same structure, the rules and precautions described above will apply equally well to either one.⁵

4. Broadening of Experience by Vicarious Contacts

5. Establishment of Desirable Reading Attitudes

The first of these aims has to do with the result of reading; the second, with the individual motivation of reading.

The teacher must seek an answer to the question: What kind of behavior is to be expected of this student as he broadens his experience by means of books; what kind of behavior is to be

⁵ See pp. 125-58 for general discussion of various test forms.

expected of a student who has established desirable reading attitudes? Answers to these questions will vary widely in accordance with the varying conditions of different schools.

Evidence that the student has progressed toward the first goal is to be found in the nature and diversity of his reading; toward the second goal, in the quality and amount of his reading. The criterion is what books, newspapers, and magazines, and how many of them, he reads habitually of his own volition.

These objectives of the English studies, then, are to be evaluated through the record of unrequired "free" reading done by the student at his leisure. They cannot be evaluated without the record, nor can the teacher guide the student to higher levels of reading without it. Records may take the form of individual booklets or diaries. Evaluation may be in terms of amount, range, and quality. Amount is measurable, and varies from school to school within very wide limits. Range is measurable when the teacher has decided what he means by range. He can make his own classifications of reading, or he may adapt the classes and subclasses used by the Committee on Home Reading in the excellent booklets prepared by the National Council of Teachers of English.⁶ Quality must be evaluated subjectively, and the teacher must be on his guard lest he be measuring himself rather than the student.

The teacher can thus devise his own rating scale for the evaluation of these objectives.⁷

6. Ability to use the Resources of Libraries

Paper and pencil tests have been designed to measure a pupil's knowledge of how information is arranged and dis-

⁶ See for other suggestions C. L. Persing and H. R. Sattley, "Discovering the Reading Interests of Maladjusted Students." *Bulletin of the American Library Association*, January, 1935, pp. 13-23.

⁷ See suggestions for similar rating scales in Chapter V.

played in books; knowledge of the contents of a dictionary; knowledge of card catalogues, catalogues of periodical literature, and of various kinds of reference books. There is evidence, however, to indicate that a good deal of such knowledge is not closely related to the ability to use it.

The teacher of English must first decide for any grade or group what skills in terms of behavior in specific situations are possible of attainment under the conditions prevailing in his own school. The supply of books in classroom libraries, the equipment of the school library, the degree of enlightenment which characterizes its administration, and the freedom allowed students in its use, all must be considered. With these in mind, the teacher can best evaluate the ability here considered through observation of how an individual pupil actually proceeds in solving a series of problems set up in terms of the material available. Successful solution of these problems will depend upon the student's knowledge of probable sources to which to refer and his ability to locate specific data in these sources.

In some schools teachers have devised cards at different high-school levels, each card containing clear directions in regard to the information desired, with a space in one corner in which is recorded the time consumed by the student in finding that information. From these records a time norm for each item is obtained for the grade level at which it is used. The student is required to return the card, together with a separate paper on which he has written the information desired and the source from which it has been obtained.

At the lower level these cards may call for ability to find information in single books, for example:

When was Samuel Pepys living?

What was the population of Marion, O., in 1930?

What is the origin of the word *Metropolis*?

At the upper level they may demand knowledge of sources and ability to use them, of any desired degree or nature; for example:

What were the general economic and political conditions in England during the reign of Elizabeth?

Graph the student population of the public high schools of the United States from 1890-1930.

The ability to use the resources of libraries must be evaluated individually. Systematically assembled information of individual progress is the material for evaluation. The system described can be evolved by every informed teacher in terms of the local conditions. It will provide the teacher with a measure of the capacities and needs of students in this field. It further will quickly identify those students who cannot use libraries because of ignorance of what libraries contain, and, when refined by the intelligent teacher, will reveal individual difficulties either in finding the desired source or in using it.

The Reproductive or Creative Objectives

1. Speaking and Writing Abilities

a. Acquisition of skill in correct language usage

As defined earlier, this phase of the major objective comprises knowledge of and habitually correct usage of the conventional tools of expression. The methods of measuring attainment of this objective are of three general types: testing on formal elements, rules, and definitions; subjective evaluation of the pupil's oral and written work; objective testing on correct usage. Each of these methods will be discussed in turn.

Testing on Formal Elements, Rules, and Definitions

Genuine mastery of formal rules and definitions will certainly facilitate correct speaking and writing; thorough understand-

ing of the meaning and uses of the various elements of the English language will assist one in employing them correctly. To belittle the importance of such knowledge is to hinder attainment of the objectives sought. Because of that importance, testing on formal elements and principles, in so far as it goes, is justified. The cause of its frequent derogation, however, is fairly obvious. To be able to state a rule or definition without knowing the meaning and import of the terms employed is useless. The pupil can derive little benefit from knowing, for example, that a period should follow a declarative sentence if he does not know what a declarative sentence is. He cannot apply the rule that a participle must have an antecedent if he is unable to recognize a participle. Understanding and application of any one rule or definition demand understanding and application of others also. Consequently, the mere ability to state rules and repeat principles is no assurance that these will function in the pupil's oral and written work. This fact does not imply that testing on definitions and rules should be neglected entirely; such testing does have a place in the program of measurement, but it should be very adequately supplemented by evaluation of the pupil's skill in usage.

One inherent weakness of most paper and pencil tests on the formal elements of English is that generally rules and definitions can be expressed clearly and succinctly in only a few ways, which quickly become stereotyped in the language of instruction. A test item expressed in hackneyed phraseology is likely to be answered on a purely mechanical memory basis. It would be better to couch the questions in original or unfamiliar terms, but fresh restatements of rules and definitions are extremely difficult to make because of the danger of ambiguities and inaccuracies. The following completion items, for example, are purely formal in nature and would undoubtedly sound familiar to all pupils.

EXAMINATIONS IN MAJOR SUBJECT FIELDS

1. The subject of a sentence is in the _____ case.
2. A pronoun must agree with its antecedent in _____, _____, and _____.
3. A word used to modify a noun or pronoun is called an _____.

To restate these items without using the words "subject," "antecedent," "modify," etc., is almost impossible. Hence such items almost invariably place a high premium on rote learning.

A more satisfactory method of testing for rules and definitions is to present the pupil with an example and require him to state or to identify the rule or definition applying to it. Objective variations of this technique have been devised which combine knowledge of principles with skill in their application. The following items are illustrative.

Directions: Each of the following sentences contains an error. In the blank after each sentence, write the word which should be substituted for the one that is incorrect. Then, in the parentheses following the blank, write the *number* of the rule (from the list at the right) which you applied in making the correction.

- | | <i>Correction</i> | <i>Rule</i> | |
|--------------------------------------|-------------------|-------------|--|
| 1. Her and I are going. _____ | (—) | 1. | Double negatives should be avoided. |
| 2. The tree shed their leaves. _____ | (—) | 2. | An adjective should not be used to modify a verb. |
| 3. We haven't no money. _____ | (—) | 3. | A pronoun must agree with its antecedent in number. |
| 4. She walks slow. _____ | (—) | 4. | The subject of a sentence is in the nominative case. |
| etc. | | | |

EXAMINATIONS IN ENGLISH

Such double testing yields truer evidence of mastery of principles, since the pupil's score on that aspect of instruction can be checked against his score on usage. Double scoring of each paper is required, however, and the two measures ordinarily should not be combined into a single score.

Understanding of the formal language elements can be satisfactorily tested by various indirect methods. The following are a few of the possible devices.

Directions: In the blank after each sentence write the case of the italicized word.

1. She gave *me* the ball. _____
2. What did *you* say? _____

This may be adapted to testing on gender, number, parts of speech, etc.

A more economical form:

Directions: Above each word in each of the following sentences write the name of the part of speech which the word represents in that sentence.

	Adj.	Noun	Verb	Adv.
Example:	The	boy	fell	down.

On sentence structure:

Directions: On the line at the left of each exercise that is a complete sentence write the letter S, and opposite each exercise that is not a complete sentence write the letter N.

- _____ 1. It is warm
- _____ 2. The concert which we attended
- _____ 3. I thought that if she wanted to go

Directions: For each sentence below indicate whether the sentence is declarative, interrogative, or imperative, by writing a *D*, *Int.*, or *Imp.* in the blank at the left.

- _____ 1. Give me the basket
- _____ 2. She asked if we would stay

Similarly for simple, complex, and compound sentences.

These few illustrations will undoubtedly suggest to the teacher other possible variations and applications.

Subjective Evaluation of Pupil Expression

It would seem that the truest behavior-evidence of the pupil's mastery over English mechanics is the degree of correctness of his actual speech and writing. From the measurement point of view, however, that evidence is not by itself a wholly satisfactory or adequate basis for appraisal. In the first place, it may not provide a fair sampling of all the skills involved in correct expression. As far as oral evidence is concerned, the teacher's opportunities for judgment are very restricted; they would be limited mainly to the pupil's classroom recitations, in which his speech may be less self-expression than mimicry of teacher and textbook. Written compositions by the pupil offer more substantial evidence, but even it may be incomplete. The pupil can readily avoid those elements and situations about which he is uncertain. If he does not know how to spell a word, he can substitute another for it. If he is not sure of a punctuation mark, he can avoid the situation by an alteration in sentence structure. It may not even occur to him to attempt those very constructions and elements which would reveal his most serious difficulties, unless he has been particularly instructed to use them.

A second difficulty in the evaluation of written work is the scoring. Unless the teacher makes an actual error count of each paper — a procedure that is extremely time-consuming — he cannot arrive at anything approximating an exact score; even then, the value given to each paper will be quite subjective, because of the differences in content, and therefore in number and variety of mechanical skills represented, from paper to paper. Strictly, the pupil's score should depend upon the *ratio* between the number of errors made and the number

of opportunities for error which his composition presented. It is obviously very difficult, however, to determine how many such opportunities there are. Uniformity in evaluating the papers for a given assignment may be furthered by the preparation of a rating scale in which a range of possible scores, from very good to very poor, is established for each of the major divisions of English mechanics. This necessitates considering each paper with reference to each of those major divisions and then adding the separate scores to determine the total evaluation of the paper, but the advantages of the method may compensate for the greater length of time consumed in its use.

These difficulties and inadequacies do not disqualify subjective evaluation of pupil expression as a means of measuring achievement, but they should be borne in mind when this type of appraisal is employed. Theme-writing, of course, should have a very important place in instruction, because it gives the pupil an opportunity to use what he has learned in a natural writing situation and it gives the teacher an opportunity to indicate the proper correction of errors that might otherwise become habitual. As a teaching device, it fulfills a unique purpose; as a measuring device, it should be supplemented by more objective methods of determining skill in correct usage.

Objective Testing on Usage

Objective tests of English usage, while a few degrees removed from the natural writing situation, compensate for that disadvantage by eliminating many of the difficulties inherent in the subjective method of evaluation just discussed. In objective examinations, a wide and representative sampling of important error situations can be included; the content is uniform for the entire group to be measured, permitting a uniform standard of rating; and the scoring is simpler, more rapid, and free from subjective factors.

Innumerable types of objective exercises have been devised for separate testing on the various elements and aspects of correct usage. Only a few can be reproduced here for illustrative purposes, but these will suggest other forms and adaptations.

Spelling

The most commonly employed type of spelling test is the list-dictation test, in which the teacher reads each word separately and the pupils write down its correct spelling. For general classroom use this has proved to be one of the most satisfactory methods of measuring spelling ability. Being a pure recall test, it more nearly approximates the natural writing situation than do the self-administering types of tests. It is simple to construct and easy to administer. The word-in-sentence list-dictation type of test, a variant form in which the teacher pronounces the word, uses it in a sentence, and then pronounces it again, reduces the possibility of errors due to misunderstanding on the part of the pupils and is essential in testing on homonyms, but is more time-consuming, both in construction and administration.

There are many kinds of self-administering spelling tests in current use in standardized examinations. Examples of six varieties (cited and evaluated by Cook ⁸) are presented below, with brief comments concerning their advantages and disadvantages.

Right-Wrong Test

Directions: If the spelling of a word is RIGHT put a circle around the R in front of it; if it is WRONG encircle the W.

- | | | | |
|---|---|----|---------|
| R | W | 1. | aranged |
| R | W | 2. | article |
| R | W | 3. | asuring |

⁸ W. W. Cook, "The Measurement of General Spelling Ability Involving Controlled Comparisons Between Techniques." University of Iowa *Studies in Education*, vol. VI, no. 6, pp. 60-84. 1932.

EXAMINATIONS IN ENGLISH

Advantages: Purely objective, easily constructed, administered, and scored; pupil must be able to choose between the given form and all other possible forms of spelling of each word, rather than between only two or a few given forms. Disadvantages: Measures only ability to recognize incorrect spellings rather than ability to spell correctly; subject to guessing.

Recognition Two-Response Test

Directions: In the following spelling list, each word is spelled in two ways. You are to select the correct spelling of each word and put its NUMBER (not the word itself) in the parentheses at the right.

- | | | |
|----------------|------------|-----|
| 1. (1) forty | (2) fourty | () |
| 2. (1) allways | (2) always | () |
| 3. (1) fought | (2) faught | () |

Advantages: Purely objective, easily scored, economical of administration time. Disadvantages: Measures recognition rather than recall; difficult to construct unless a list of frequent misspellings is available; offers only two choices in each item; subject to guessing.

Recognition Four-Response Test

Directions: Each of the following words is spelled in four ways. You are to select the correct spelling of each word and put its NUMBER (not the word itself) in the parentheses at the right.

- | | | | | |
|---------------|------------|-------------|-------------|-----|
| 1. (1) fourty | (2) forty | (3) fortey | (4) fourtey | () |
| 2. (1) always | (2) alway | (3) allways | (4) alaway | () |
| 3. (1) fot | (2) fought | (3) foght | (4) faut | () |

Advantages: Purely objective, easily scored, reasonably economical of administration time; offers sufficient choices to reduce factor of guessing. Disadvantages: Measures recognition rather than recall; construction requires recourse

EXAMINATIONS IN MAJOR SUBJECT FIELDS

to a list of common misspellings; subject to guessing to a certain extent.

Column Proof-Reading Recall Test

Directions: Some of the words in the following list are spelled incorrectly. If a word is spelled correctly, place a C on the line opposite it. If it is spelled incorrectly, write the correct spelling of the word on the line opposite it.

1. aranged _____
2. article _____
3. asuring _____

Advantages: Easy to construct; administration and scoring time reasonably economical; tests both recognition and recall of correct spellings. Disadvantages: Not strictly objective; open to guessing to some degree; list of frequent misspellings needed for its construction.

Word-in-Sentence Recall Test

Directions: Write the correct spelling of the underlined word in each of the following sentences.

1. It was an eskwizit piece of lace. _____
2. I am greatfull for your assistance. _____
3. The accident was unfortunitate. _____

The underlined words are deliberately mutilated as much as possible without destroying their phonic individuality, in order to avoid suggesting any part of the correct spelling to the pupil. The pupil is expected to recognize the intended word by the context and the phonetic quality of the misspelled word, and to write its correct spelling in the blank.

Advantages: Strictly a recall form of test; reduces guessing to a minimum. Disadvantages: Not strictly objective; requires considerable time for construction, administration, and scoring; the gross misspellings suggested may influence the

pupil's spelling of the word; because of mutilation, the words may measure intelligence and comprehension rather than spelling ability.

Sentence Proof-Reading Recall Test

Directions: Many of the following sentences, but not all of them, contain misspelled words. You are to find these misspelled words, underline them, and write the correct spelling in the space to the right of each sentence. If all of the words in a sentence are spelled correctly, place a C in the space.

1. Place the picture on the bulliten board. _____
2. Ordinarily I work ten hours per day. _____
3. I have no appology to make. _____

Advantages: Principally a recall test; factor of guessing almost eliminated. Disadvantages: Not strictly objective; requires considerable time to construct, administer, and score; irrelevant factor of proof-reading ability may influence score.

On the basis of his experiment, Cook concluded that the proof-reading recall test was, in general, superior to the other five types illustrated above.

Objective devices such as these need be resorted to only in special cases where the list-dictation method is less feasible. They are useful in the separate testing of several different groups when strict comparability of results from group to group is desired, in any situation where variations in dictation and irrelevant difficulties due to enunciation are to be avoided, and in the testing of pupils whose hearing is defective. In most general classroom testing, the list-dictation method is satisfactory and preferable.

Punctuation

Perhaps the best type of punctuation test, because most closely analogous to the "free" writing situation, is the dicta-

tion exercise. The teacher dictates sentences or paragraphs which the pupil copies and punctuates properly. The administrator must avoid giving hints to the pupil by the way in which the material is dictated; for example, suggesting commas by pauses, periods by falling inflection, interrogation points by rising inflection. Because the material must be read slowly and because the pupil may have to recopy it in order to submit a legible paper, the time required for dictation tests is considerably greater than for other types of examinations. The scoring also is time-consuming, and may present difficulties due to illegibility of writing, to the fact that the pupils often introduce errors not previously anticipated by the teacher, and so on. When properly administered, however, the dictation test has the value of permitting the pupil to demonstrate his knowledge without external suggestion, and of making somewhat the same demands upon the pupil as would the writing of a theme, while maintaining for the teacher a fair uniformity in the range and distribution of error situations. It also facilitates the detection and correction of the punctuation difficulties of individual pupils. As a teaching device, it has similar advantages, which need not be elucidated at length here.

A fairly good substitute for the dictation test is that in which the pupil is presented with an unpunctuated paragraph or list of sentences and told to provide the correct punctuation. One danger in this form is that the complete lack of punctuation may confuse the meaning of the material and thus lead to unintentional errors on the part of the pupil. If punctuation containing many errors is already included in the test material and the pupil is told to make the necessary corrections, the exercise will measure both recognition and recall of correct usage.

The scoring of such usage tests is likely to be very difficult

and partially subjective unless special provision for objective control is made. If each line of the material contains only one "planted" error, the pupil can be instructed to make the correction by writing in a blank at the right of the line the word or words adjoining the error and indicating the punctuation which properly should precede or follow them. This method, which has been used in the Iowa Every-Pupil and the Cooperative English correctness tests, permits rapid and efficient scoring by means of a scoring key. The key, of course, must include, not only the most common form of correction of each error, but also any acceptable alternate forms; for this reason, the error situations included in the test material must be clear-cut and not open to argument on the grounds of deviation of modern usage from the traditional forms taught in the older textbooks and courses of study.

A scoring device which enables the pupil to make his corrections directly in the text of the test material and which permits objective scoring was developed in the construction of the 1935 Iowa Every-Pupil Test of Basic Language Skills for Grades 6, 7, and 8. A part of the first page of this test and the corresponding part of the scoring key are reproduced on pages 422-423. The scoring key itself is a duplicate of the corresponding sections of the test, with the proper punctuation marks indicated in large size and heavy inking so as to be readily discernible by the scorer. The blank portion of the key to the right of the dotted line is cut away, so that, when ready for use, the key has a jagged right-hand margin which parallels the position of the errors in the test itself. When the key is laid upon the test page, both the error situation and the proper correction, side by side, are plainly visible to the scorer, and each line can readily be marked as right or wrong in the margin of the test page. Here again, only one error situation can be included in each line.

PART III — PUNCTUATION

DIRECTIONS: Every one of the sentences below contains *one* error in punctuation. You are to correct each error, making your correction right in the sentence itself. The samples below will show you how the sentences should look after you have made the corrections.

If a punctuation mark is missing, put it in where it belongs. (In some of the sentences you may need to put in quotation marks at *two* places in the sentence. All other sentences contain an error at only *one* place.) For example, in the first sample the quotation marks were originally left out around "Home Sweet Home."

If the wrong punctuation mark has been used, draw a vertical line through it and then put the correct punctuation mark beside it. For example, in the second sample the question mark at the end of the sentence is incorrect. It has therefore been crossed out and a period placed beside it.

If a punctuation mark is found where none belongs, simply draw a line through it. For example, in the third sample a comma does not belong after "town." It has therefore been crossed out.

In the fourth sample, the apostrophe was originally left out of "doesn't." It has therefore been written in. In the fifth sample, the period after "me" is incorrect, so it has been crossed out, and a comma has been written beside it. Notice that the comma is written *inside* the quotation marks.

Correct the rest of the sentences in the same way. Make your corrections very plainly, and draw your vertical lines exactly *through* any wrong punctuation marks.

Sample 1: He played "Home Sweet Home."

Sample 2: I wonder who he is?

Sample 3: I went to town/ yesterday.

Sample 4: It doesn't look like rain.

Sample 5: "Give me," he said, "some cake."

1. Who can work this problem (1)
2. We drove through Mt Vernon this summer. (2)
3. The rain and sun made the gardens grow (3)
4. When the time came to go nobody was ready. (4)
5. "Fire" screamed the boy, as he ran from the house. (5)
6. Would you like to come too? (6)
7. We left at ten oclock. (7)
8. Nevertheless they wished to go. (8)
9. Her address is 822 W Ash Street. (9)
10. At one time he lived in St Paul. (10)

SCORING KEY FOR TEST C, PART III

DIRECTIONS: This page constitutes the scoring key for page 6 (the first page of the Punctuation Test). Corrections of all errors in punctuation are clearly indicated, since they have been written in by hand in heavy inking and extra large size.

With a pair of scissors or a sharp pen knife, cut very carefully along the irregular dotted lines indicated by the arrows. The key will then be ready for use. To score column 1 (items 1 to 10) of the pupil's paper, proceed as follows. Lay the key on the pupil's paper so that each sentence on the key is directly over the corresponding sentence of the pupil's paper. Then move the key horizontally about an inch to the left. Each of the proper corrections made by the pupil will then appear about an inch to the right of the corresponding correction on the key. With a colored pencil, draw a straight horizontal line through the small number in parentheses after each sentence that has been properly corrected. If the correction on the pupil's paper does not correspond with that on the answer key, draw a cross through the number in parentheses at the end of the sentence. Disregard completely any corrections that the pupil may have made at other places in the sentence than that indicated by the heavy corrections on the key.

After having scored column 1, turn the key around and place the other irregular margin over column 2, adjusting the key as before so that the corrections on the key are just opposite and about an inch to the left of the corresponding places on the pupil's paper. Score each item as before by drawing a horizontal line or a cross through the number in parentheses to the right of the item.

Prepare the key for page 7 in the same fashion, and score the items in the same way.

The total score on Part III is simply the number of sentences that have been properly corrected, and can be determined by counting the number of horizontal lines that have been drawn through the numbers to the right of each column.

Part III - Punctuation
Page 6, Col. 1

1. Who can work this problem?
2. We drove through Mt. Ve
summer.
3. The rain a
grow.
4. When the time came to go, no
was ready.
5. "Fire!" s
from the house.
6. Would you like to come, to
7. We left at ten o'clock
8. Nevertheless, th
9. Her address is 823 W. As
10. At one time he lived in St. Pa

(Cut along the dotted line indicated by arrows.)

EXAMINATIONS IN MAJOR SUBJECT FIELDS

In an effort to achieve strictly objective scoring, the constructors of standardized tests have sometimes employed the form of exercise in which the pupil indicates merely whether a given sentence is punctuated correctly or incorrectly. The evidence yielded by such an exercise is very inadequate. The pupil may properly mark a sentence as right or wrong because he "feels" it is so, without knowing why or in what specific instance it is right or wrong. Even in those variations which require him to indicate the type of error by encircling a letter (as the letter *c* for comma errors) in the margin, there is no means of knowing whether he detected the true error or whether he had reference to an imaginary mistake. At best, the exercise measures only the pupil's ability to recognize the correctness of the given punctuation, and may neglect entirely his ability to supply the proper punctuation where needed. In any case, it is open to guessing.

Capitalization

Here again, the dictation exercise has the same general merits and demerits as in testing on punctuation, and can be used to advantage in the ordinary classroom situation.

Most of the strictly objective devices for measuring skill in capitalization have the same fundamental structure: the presentation of sentence material in which the pupil is to capitalize the proper words. The variations of this technique are principally differences in the manner in which the pupil indicates the correct responses. Underlining of the words to be capitalized is sometimes employed, but this method is clumsy to score and does not permit the inclusion of improperly capitalized words in the test material. Both of these disadvantages can be overcome by requiring the pupil to write correctly, in a marginal blank, the word that contains a capitalization error, or, as in the Iowa Every-Pupil Test of Basic Language Skills,

by numbering each word in the line and requiring the pupil to indicate the number of the incorrect word or to indicate a zero if the line contains no errors. In the latter two techniques, of course, only one real error can be included in a line; to the poorly equipped pupil, however, every important word in the line will present a possible error-situation, and the amount of discrimination required in selecting the correct response is therefore much greater than may at first be apparent. This fact should be borne in mind in constructing the test material.

Grammatical Usage

The more commonly employed techniques of testing for grammar or language usage are illustrated in the following examples.

He (doesn't don't) want to go home.

The pupil may be directed either to underline the correct word or to cross out the incorrect word. Both methods are difficult to score. The scoring may be simplified by the following device.

- Directions: In each of the following exercises, only one of the two words or phrases printed in bold-face type is correct. You are to select the correct word or phrase and write its *number* in the parentheses at the right.

He (1) **doesn't** (2) **don't** want to go home. ()

These are straight two-choice recognition tests. A much better form — probably the most satisfactory method of testing for grammatical usage — is illustrated by the following items, which test both for recognition of errors and for recall of correct usage.

Directions: Some of the following sentences contain errors in grammar; others are entirely correct. Read each sentence carefully. If it contains a word that is used incorrectly, write

the correct form of that word in the blank at the right. If you think the sentence is correct as it stands, write a *C* in the blank.

He don't want to go home.

They brought gifts for John and me.

Attempts have been made to test for pure recall by omitting the test word and requiring the pupil to supply the proper form.

Each of the boys took own books home.

In this instance the pupil is expected to supply the word "his." Exercises of this type are extremely difficult to build because of the danger of ambiguity. The intended word must be made obvious by the context, and there must be no possibility of satisfactory alternatives. The number of error-situations to which the exercise is adaptable is limited by these requirements. Also, it measures the irrelevant factor of comprehension ability to a large extent.

Tests in which the pupil merely marks a sentence as right or wrong have virtually no usefulness for the classroom teacher, since they measure only skill in detection and ignore the much more important correction ability. Furthermore, they too readily permit guessing.

In considering achievement tests which purport to test grammatical correctness and correct usage of idiom and phrase, the teacher should have in mind that weaknesses in achievement tests sometimes are incident to weaknesses in the curriculum. Though teachers and published descriptions of courses agree that what is to be taught is functional grammar, there is no exact knowledge and certainly no agreement as to what functional grammar is.⁹ Achievement tests in English usage

⁹ For promising explorations of the field see H. N. Rivlin, *Functional Grammar*. New York: Bureau of Publications, Teachers College, Columbia University, 1930. Also Verna L. Newsome, "Making English Grammar Function," *English Journal*, January, 1934, pp. 48-57.

will be still more effective when there is greater knowledge of the relative importance of different items, and of the relative difficulty of items, and when teachers themselves come to some agreement as to what grammar is functional. A detailed analysis of the grammar content in 22 junior and 22 senior high-school courses shows that the amount of grammar which is considered functional varies from 45 topics to 149. The report of the nation-wide study for the rebuilding of the English usage curriculum shows wide differences between the estimates of relative difficulty of types and items of English usage, and the actual relative difficulties found as the result of investigation.²⁰ More effective tests will naturally result from more certain knowledge of what a rational curriculum of English usage is, and of the degree of difficulty of the items which it includes. Racial, geographical, and economic factors unite to make the needs of students extremely varied in regard to the English usage curriculum. The needs of two different schools in the same town may be widely diverse. Hence the teacher of English may very profitably collect material, test its different items for difficulty, and arrange them in test form for the needs of the local situation. In this task he will apply the principles of test-making, and study the form of available tests, for some of these have proved to be extremely effective agents in the establishment and maintenance of skills, a fact which is evidenced by the scores made in the state of Iowa when compared with the norms of other public school systems as measured by the same instruments. Strong evidence is also adduced in the progressive improvement in test scores from grades 3 through 12 which has resulted from the emphasis which schools co-operating in the O'Rourke study have placed upon the elimination of certain errors.

²⁰ L. J. O'Rourke, *Rebuilding the English-Usage Curriculum to Insure Greater Mastery of Essentials*. Washington, D.C.: The Psychological Institute, 1934.

Amid the wide divergence of theory and practice in various parts of the country in regard to what grammar is to be taught, how may the teacher of English evaluate the results he is accomplishing? In Appendix A at the end of O'Rourke's report already referred to, is a list of 79 items of usage arranged in order of importance and in the order in which they should be presented. The teacher has in this list a basis for what is to be tested from the point of view of a nation-wide survey. But it should be pointed out again that local conditions vary widely. In some school populations, habitually correct usage of a considerable number of these items has been established in the home. The teacher of English will select the test indicated after the status of his group in habitual English usage is known to him, and he will use tests of grammatical correctness as much to discover what is already known as to determine attainments in terms of national norms.

Sentence Structure

Because of the mechanical difficulties involved in their preparation, administration, and scoring, the construction of objective examinations is less practicable in testing on sentence structure than in testing on other phases of English mechanics. The simplest type of exercise, which is the right-wrong technique, is inadequate for reasons stated in preceding paragraphs. The type in which the pupil selects from a group of sentences the one which contains no structural defects yields a more satisfactory measure of detection ability and is relatively easy to construct. For classroom usage, it is perhaps the best of the objective techniques.

More complex forms, which attempt to measure correction ability, have been developed for use in some standardized tests. For example:

(1) of maple sugar, (2) enjoyed most, (3) is
the making, (4) of farming, (5) the part, (6)
by a boy

5 4 2 6 3 1

Here the pupil reorganizes the sentence and indicates the result by writing the numbers of the words and phrases in proper sequence. Besides being difficult to construct, such forms consume an inordinate amount of administration time. They are likely to be very confusing to the pupil, and ambiguities may operate to the detriment of his score. The exercise may measure general intelligence and comprehension to as great a degree as it does ability in handling sentence structure.

Vocabulary

The most common form of vocabulary test is that in which the pupil is given a list of words, opposite each of which is given a number of words (usually four or five). In each item the pupil is required to select from these words the one which most nearly approximates the meaning of the test word and to write its number in a blank or in parentheses. For example:

Nauseate: (1) sadden (2) worry (3) anger (4) excite
(5) sicken..... (5)

In such an item, the response words should not themselves be more difficult than the test word. The wrong responses should be such as will appear plausible to the uninformed pupil. Words that are similar to the test word in either sound or appearance, but not in meaning, will make good wrong responses, since they will attract the pupil whose knowledge is uncertain and who may have a tendency to guess. All the responses should be as nearly homogeneous as possible; i.e., one should not construct an item in which one response is a noun, another a verb, another a prepositional phrase, etc.

The variant form of item in which the test word is presented

in a sentence may be useful in measuring the pupil's understanding of words that have more than one important meaning, and for which the context is necessary to indicate which of the meanings is involved. This type of item requires even greater care in its construction than does the type exemplified above, since the wrong responses must appear plausible with reference to the context but, at the same time, must not be defensible as the correct response.

It should be noted that vocabulary tests of the types just considered are strictly valid only for the measurement of the pupil's passive or reading vocabulary and do not indicate reliably the extent of his active vocabulary, i.e., the vocabulary which he uses habitually in his own speech and writing. It should be noted also that there is a real distinction between speaking and writing vocabularies, and that both may not be effectively measured by the same instrument. Valid measures of the latter types of vocabularies can be obtained only through an analysis of the pupil's own speech and writing. Such analyses, obviously, are extremely difficult to make. About the best that the limited time of the teacher will allow is a subjective evaluation of the pupil's spoken and written language from this point of view.

Combined Objective Tests of Usage

There are two principal methods of testing simultaneously the pupil's skill in all the phases of English usage: the dictation exercise and the proof-reading test.

In the dictation exercise, the pupil copies verbatim the material dictated to him and then recopies it, making all necessary corrections in grammar, supplying the correct punctuation, capitalization, spelling, etc. The method requires a considerable amount of time in both administration and scoring, and, as has already been pointed out, the scoring difficulties are ap-

preciable. Where the teacher can devote to it the time and effort demanded, however, this type of testing is desirable because of its resemblance to the "free" writing situation. In order to simplify the scoring, the pupil should be instructed to make as few changes as possible in the actual text of the material. Grammar errors should be mainly one- or two-word situations; if changes in entire sentence structure or in organization are required, the task of appraisal will be little different from subjective evaluation of an original theme.

The general merits and limitations of proof-reading tests have already been described or implied in the preceding sections. The mixed-error type of proof-reading test has the particular advantage that, as in his own writing, the pupil must be alert to all phases of English mechanics simultaneously. The Iowa Every-Pupil Test in English Correctness is an example of this type of test. The examination consists of six short themes containing errors of spelling, punctuation, grammar, etc. No line includes more than one error, and some lines are entirely correct. The pupil is instructed to provide the correction for the error in a blank to the right of the line, or to write an "R" in the blank if the line is correct.

Certain weaknesses of this kind of test must be recognized. The irrelevant factor of sheer proof-reading ability may, of course, operate to some extent. The sampling of errors represented in the test must exclude some which may be important but which cannot be economically corrected in the manner directed or cannot be objectively scored. There is some danger that, knowing each line contains only one error, the pupil may "correct" an imaginary error and carelessly overlook the real or "planted" error in another part of the line. For this reason, and also because the segregation of types of errors cannot be handled very conveniently in scoring, the mixed-error proof-reading test has little diagnostic value. For the measurement

of general achievement in the use of English mechanics, however, it is one of the most practicable methods at present available.

A comparatively new test form in this field is that called the "construction shift," which was used in the Wisconsin Language Test — Gamma, and is now used in the Cooperative English Test, Series 1. In a recent article,¹¹ this testing technique has been described as follows: "The form of the test is simple. It is not a purely objective test; it allows some variation in correct response. It requires actual writing by the student. The student is instructed to make a certain specified change in a given sentence (which may be correct, incorrect, or merely poor), together with additional changes made necessary by the specified change. For example, the student is told to 'substitute *a few* for *one*' in the sentence, 'There was one of the men waiting for me.' The student is to make the smallest number of changes possible, and to make no changes in meaning, tense, word order, etc., except those necessitated by the specific directions. In the example given above, the student must write, 'There *were* a few . . .' (not 'There *would have been*' or 'There *are*'). The student need not rewrite the entire sentence, but may make the necessary changes by crossing out and adding words. In the example given the student should cross out 'was one' and write in 'were a few.'"

The following are a few of the examples cited in the article to illustrate the use of this technique in measuring knowledge of various points of usage.

Punctuation:

Change *but* to *nevertheless*: You cannot go to the concert, but you can read about it in the newspaper.

..... concert; nevertheless, you (OR) concert. Nevertheless

¹¹ John M. and Ruth C. Stalnaker, "A 'Construction Shift' English Test." *English Journal* (College Edition), vol. xxiv, no. 8, October, 1935.

Grammar:

Subordinate the first clause: I did not approve of what he said, and I told him how I felt.

Because I did not approve of what he said, I told him how I felt.

Start the sentence *Upon graduating from school*: My father took me to Europe.

Upon graduating from school, I was taken to Europe by my father.

Change *suspect* to *suspicion*: I suspect that she is guilty.

My suspicion is that she is guilty. (Not I *suspicion* that)

Capitalization:

Omit the word Amherst: Three of my friends went to Amherst College.

..... went to college.

General emphasis and clarity:

Change the sentence so as to emphasize the fact that Homer was the maker of a nation: Homer was the maker not only of a nation but also of a language.

Homer was the maker not only of a language but also of a nation.

. The test is not purely objective, since some of the sentences can be changed in more than one correct manner, but this difficulty can be met by anticipating such variations or by observing and rating them consistently in the scoring. Care must be taken in the construction of the items to reduce to a minimum the possibility of variations in response. Users of this testing technique have recommended it highly on the grounds of reliability of scoring, adaptability to varying levels of difficulty, and apparent relationship to general writing ability. It appears, however, to require considerable ingenuity on the part of the test constructor, particularly if one is to avoid the danger of placing an undue premium upon general intelligence or upon the ability to comprehend directions.

b. Power of Expression

The ability to organize and express effectively the results of the individual's own thought and expression has been included in the tabulation of objectives because it is a generally recognized goal of courses in composition. It is, however, an example of those desirable outcomes for which no adequate measuring instruments have yet been devised. The subjective factors involved in evaluating an individual's composition cannot readily be controlled. While there may be loose agreement upon what elements, in general, tend to produce quality in writing, there is likely to be far less agreement in the evaluation of any specific composition. For this reason, few concrete suggestions can be given for the measurement of the ability here considered.

Composition scales have been developed which purport to provide the teacher with a standard for rating the themes of his pupils. These scales generally consist of a set of short sample themes, ranging from very poor to very good, with a numerical value assigned to each. The teacher compares a particular theme of one of his own pupils to this set, decides which sample is most nearly comparable in quality, and gives to the theme the numerical value stated for the sample in the scale. Such scales have their principal value in the lower grades, where the theme topics assigned to the pupils are simple and the general composition work quite elementary. They do not constitute a truly objective standard, since the scorer exercises subjective judgment in comparing any one theme to the scale. Furthermore, the greater the divergence between the nature of the pupil's compositions and the nature of the samples in the scale, the greater will be the influence of this subjective factor and the less the value of the scale used. The teacher may find standardized composition scales of some usefulness as an external check upon his own individual scale of

rating, but at the upper levels they will have little other utility. At these levels, the teacher must depend mainly upon his own judgment, basing his appraisal upon consideration of such elements as soundness of ideas presented in the theme; clarity and coherence of organization; selection, arrangement, and emphasis of details; and vividness and convincingness of expression.

If a somewhat impersonal topic is assigned to the entire class, the teacher can, to a certain extent, determine beforehand what may be expected in a good theme on that topic and may evaluate the themes in relation to that subjective norm. The predetermined standard, however, must be sufficiently flexible to allow for worthy methods of presentation and treatment that have not been anticipated. Also, the topic assigned must not be concerned with a subject that has already been thought through or discussed in the classroom or in the study materials used in the course; it must demand from the pupil originality of thought.

If the topic is a more personal one, it may be helpful to delay the actual grading until all papers have been read and arranged in order of general merit. With the range of talent thus defined, the general grade limits may be more readily determined and each paper may be more equitably graded with reference to the others.

2. Development of Desirable Attitudes

No specific definition of the desirable attitudes to be developed through courses in English has ever been agreed upon. Roughly, perhaps, they may be described as, on the one hand, the desire to speak and write correctly, and, on the other, interest in utilizing acquired knowledge and ability in self-expression wherever opportunity occurs. The former will, of course, be evidenced indirectly through the pupil's general

test performance and classwork. For evidence of the latter, the teacher must depend upon personal observation of the behavior of each individual pupil in those situations which afford opportunity for expression.

With reference to speaking attitudes, the teacher might consider such questions as: Does the student participate of his own volition in class discussions and in meetings of school organization; does he take his share in club programs; does he volunteer for programs for assemblies, and for a speaking part in dramatics; does he participate in debates formal or informal; does he share in conversation at social gatherings?

In so far as writing is concerned, the points to be considered might include the following: Does the student do more than the minimum when writing opportunities occur in the classroom; does he prepare of his own accord written contributions to class discussions; is he a member of the writers' club; does he contribute to school publications; is he interested in writing items of school or other news for the local papers; does he produce stories, poems, etc., for his own satisfaction?

Evidence of the development of these attitudes cannot be evaluated if it is not recorded. Either in the teacher's own record or preferably in the cumulative record of the individual student, brief statements should be made of acts of behavior which for any particular school are deemed most indicative; the significance and nature of such acts will depend on the opportunities provided for the exercise of them. As these records are kept and evaluations attempted from them, the growth of desirable attitudes in speaking and writing may be found to be associated with typical patterns of behavior in particular schools; and these typical patterns may suggest short-cut methods to a more rapid but still valid evaluation.

The nebulousity of any objective that concerns "attitudes," the difficulties of culling evidence of their development, and

the very dubious significance of whatever evidence is collected throw serious doubt upon the wisdom of expending upon the measurement of that objective extensive time and effort which might be employed more profitably in other directions. Perhaps, after all, the pupil's general achievement in the field of English, when he is effectively guided therein, both proceeds from and leads to desirable "attitudes," and is therefore the best indication of their possession.

QUESTIONS FOR DISCUSSION

1. Discuss the relationship between the problems of instructional objectives and testing in the field of literature. Analyze independently what you consider to be the desired objectives of instruction and indicate which of these you believe can be measured adequately. Where possible, cite actual instances — culled either from your own experience in test construction or from standardized tests — to support your contentions.
2. What special problems must be considered in testing reading comprehension with reference to literature? Supplement as extensively as you can the points suggested in the chapter. Indicate any points which you consider important but which you believe are not amenable to measurement, stating reasons for your opinions.
3. Discuss the advantages and disadvantages of the "essay" type of question in measuring comprehension and appreciation of literature. Suggest specific methods which may be employed in this type of measurement. Present, with a critical evaluation, examples of such variations in approach as have already come to your attention.
4. Present examples of various types of objective exercises employed in testing comprehension of literature. Analyze their merits and shortcomings. Which types do you regard as superior, and why?
5. Select three literary passages and construct objective multiple-choice questions for each passage. Evaluate the extent to

which your exercises cover the various reading problems peculiar to comprehension of literature.

6. What are some of the structural defects to be particularly avoided in the construction of objective exercises of the question-on-passage type? Find, in published or teacher-made tests, examples of items defective in these respects.
7. What factors should be considered in the selection of material for this type of exercise?
8. How can superficial or rote learning be penalized in testing for acquaintance with literature? Construct 10 multiple-choice items representing different methods of approach.
9. For each of three literary selections, construct three multiple-choice items each of which tests for progressively greater depth of familiarity with the selection.
10. Find six examples of items, purporting to test for acquaintance with literature, which contain serious clues or cues to the correct response.
11. Construct 10 multiple-choice items representing a variety of approaches in testing for knowledge of literary history.
12. Construct at least three different *types* of objective test exercises for each of the major objectives of instruction in literature. In each case, compare the three types and state the relative merits of each type.
13. Construct a comprehensive objective achievement test for a particular course in literature, in whatever form you prefer. Submit also a critical analysis of this test, from the points of view of content and form.
14. Construct an "essay" examination over the same material, and compare the two tests, noting in what respects each is superior and inferior.
15. Suggest and discuss methods of obtaining an approximate measure of achievement with reference to the less tangible and less immediate objectives of instruction in literature.
16. Why does testing on rules and definitions yield an inadequate measure of achievement in English mechanics? What structural weaknesses are common to this type of test?

EXAMINATIONS IN ENGLISH

17. Select, from available tests, four examples of different methods of testing for *application* of grammatical rules and definitions. Evaluate each.
18. What objections may be raised to basing an evaluation of achievement upon the pupil's own written work? What defenses are tenable?
19. Does the dictation method meet the objections referred to in the preceding question? How? What are its limitations?
20. What are the major advantages of objective tests of achievement in English mechanics?
21. What type of spelling test is most practicable for general classroom use? Why?
22. Find examples of four self-administering types of spelling tests, and evaluate each. What factors should be considered in determining the type of test to be used?
23. Construct a proof-reading exercise on punctuation in which special provision for the pupil's corrections is made apart from the textual matter. One in which the pupil's corrections are made in the text itself. What advantages does either method have over the other?
24. Find and evaluate examples of several different techniques for measuring capitalization ability.
25. Which is preferable, the recognition or the recall test, as a measure of grammatical usage? Why? Construct an exercise which you believe embodies the desirable characteristics of a good test of grammatical usage.
26. What difficulties hinder the construction of objective tests on sentence structure? Find and evaluate three illustrations of various methods of testing this phase of English correctness.
27. Construct 10 multiple-choice items on vocabulary. What factors should be considered in building such a test?
28. How do differences in speaking and writing vocabularies affect the validity of a vocabulary test?
29. What is the principal advantage of tests which measure simultaneously all phases of English usage?

EXAMINATIONS IN MAJOR SUBJECT FIELDS

30. What are the major limitations of proof-reading tests of English usage? Construct a one-page test of this type.
31. In what ways is the "construction shift" test superior or inferior to the proof-reading test?
32. Discuss the obstacles to measurement of ability in composition. Suggest any methods of measurement with which you may be familiar.
33. Discuss the relationship between test construction and the measurement of objectives concerned with "attitudes."

PART III

THE FUNCTIONS AND LIMITATIONS OF
EXAMINATIONS

CHAPTER IX

THE USES AND ABUSES OF EXAMINATIONS

1. The Need of a Philosophy of Examinations

PREVIOUS chapters of this book have described the modern testing movement, analyzed the principles and procedures of test construction, discussed detailed objectives, and examined the special testing problems of the five major subject-matter fields. It may be worthwhile to conclude with a more general, non-technical discussion of a fundamental question which must have occurred to many readers, namely, the always persistent (and perilous) question, What's the use? What do we accomplish by all this testing and examining anyway? Is it worth all the effort and money it costs? Do we perchance do harm instead of good, or harm as well as good, with our examinations, and especially through the uses we make of their results?

. In short, it may seem that we need, not only techniques, but also some philosophy of tests and examinations, dealing especially with the right uses of such instruments and their wrong uses or abuses.

It is the purpose of this chapter to attempt such a philosophy; to formulate, tentatively at least, some general doctrine as to the right use of examinations of all kinds, old-type and new-type alike.

To this end I shall review the various uses which we educators do actually make of examinations and tests, with such appraisal of these uses as might naturally occur to any experienced teacher or administrator who gave critical thought to the matter. Perhaps when we have finished such a survey

we shall see more clearly both what we have been doing in the past and what we ought to do.

2. Kinds of Examinations

It may be well to begin by summarizing the kinds of examinations we now have to confront. Classifying by what it is sought to measure in each case, we find at least five kinds:

1. *Achievement tests*. These include: (a) the old *essay-type examinations*; (b) *problem-type examinations*, used from of old in courses in mathematics and some other subjects; (c) *oral examinations*, which have long been in vogue in connection with doctoral dissertations and have recently come to play a part in the Comprehensive Examinations for Honors at Swarthmore and similar comprehensive examinations now required for honors or for graduation at many other colleges; and (d) *objective achievement tests* (new-type), of which the tests prepared by the Cooperative Test Service and those used in the Iowa Every-Pupil Testing Program may serve as examples.

2. *General intelligence tests*, such as the Binet-Stanford and Otis tests.

3. *Aptitude tests*: the Scholastic Aptitude Test given by the College Entrance Examination Board, the American Council on Education Scholastic Aptitude Test, the Seashore Measures of Musical Talent, etc.

4. *Interest tests*, of which the Strong Vocational Interest Test is probably the best known.

5. *Personality inventories and ratings*, of many kinds, as different as the Bernreuter Personality Inventory and the Trait Study recently devised by Dr. Eugene Randolph Smith's committee, working under the Progressive Education Association's Commission on the Relation of School and College.

It will be seen at once that of these five kinds of examina-

tions the last four are new, having made their way into schools and colleges within the last twenty years. For these four kinds it can hardly be said that we have as yet found any general, commonly accepted, institutionalized use or uses. So far, most of us have handled these tests and their results very informally; the scores or rankings have served as clues or cues for personnel officers in conferences with advisees, but have been entered on official records only as *obiter dicta*, if at all. This situation is probably due in part to the fact that we have hardly had time as yet to assimilate these new measures into our old routines, but partly also to the nature of these tests and their results, which, happily, do not lend themselves readily to the kind of uses we have been accustomed to make of examinations.¹

It follows that my proposed survey of the uses of examinations will be concerned almost wholly with the first kind, namely, achievement tests. And under achievement tests we shall really be dealing chiefly with the essay-type, since we may safely guess that ninety per cent of our actual examining still employs that type; with some help from problem-type and oral examinations (which are only auxiliary variants of the essay-type), and with still limited but rapidly growing use of the new-type objective achievement tests.

3. Actual Uses of Examinations

When one begins to meditate upon the achievement tests, old and new, which are actually given in schools and colleges, one can hardly fail to be astonished by their multiplicity; by

¹ The beginnings of a general, institutional use of aptitude test results may be found in the regular figuring of such results into a general predictive index — "Bogie" at Princeton, "General Prediction" at Yale, "College Aptitude Rating" at Minnesota. See A. B. Crawford, "Aptitude Testing in Personnel Procedure," *Bulletin of the American Association of Collegiate Registrars*, July, 1934; and J. B. Johnston, "Advising College Students," *Journal of Higher Education*, June, 1930.

the sheer quantity of examining in which we are engaged. The following list makes no attempt to be exhaustive: daily quizzes, weekly quizzes, and monthly quizzes, at all levels; semester examinations and year examinations, at all levels; college entrance examinations; comprehensive examinations for promotion to the senior college and for graduation or honors; master's and doctoral examinations; and professional licensing examinations (for lawyers, teachers, physicians, dentists, pharmacists, nurses, opticians, veterinarians, mid-wives, barbers, stationary engineers, etc.). Again we are impelled to ask, why do we give such countless tests?

Probably many persons will answer immediately that the obvious and legitimate purpose of practically all this achievement testing is the *maintenance of standards*; which seems to mean either one or both of two things: the imposition and enforcement of a prescribed curriculum; or the enforcement of some minimum degree of attainment.

Upon reflection it is likely to be added that a second, closely related purpose is *selection*, involving also rejection; the segregation of the fit from the unfit, the sheep from the goats, for some particular purpose.

And in the literature on the subject we find still other purposes cited as supplementary or in some cases primary. It is said that examinations:

Provide a powerful *incentive to study*;
 Constitute a *method of instruction*;
 Stimulate or even enforce *improvement of teaching*;
 Afford a basis for the *appraisal of teachers and departments*;
 May be of assistance in the *accrediting of schools and colleges*;
 Furnish data for *educational guidance*;
 And accumulate *materials for research*.

This is certainly a notable miscellany of objectives; many of them possessing obviously very great social importance. We

are fortunate indeed if all or any of them can be adequately served by series of measurements the great majority of which are of unknown and highly dubious reliability. But let us consider these several purposes or uses one by one with such comments as may seem appropriate.

4. Standards-Enforcement

First, the *maintenance of standards*, including, as already noted, the enforcement both of prescribed subject matter and of some more or less definitely envisaged degree of attainment.

If one is to raise any objections here, he must tread softly, because he is approaching what is to many educators in service, especially many of the older ones, the Ark of the Covenant. When those of us who are now in our forties and fifties were learning our trade, "Standards" was the great word, the new gospel, in American education. To set Standards, and enforce Standards, and raise Standards, and raise them ever more, was nearly the whole duty of teachers and principals and presidents. Let me confess that I learned that gospel in my first job, from men who were leaders in their generation, and that for twenty years I never doubted that it contained practically everything needful for educational salvation.

And please note that it was a real and salutary gospel in its day. For American education in the Nineties was a variegated hodgepodge of uncoordinated practices, which had never undergone any scrutiny from anyone, and many of which were shoddy, futile, and absurd beyond anything we now conceive of; and the Age of Standards, as the period from 1890 to 1915 may come to be called, brought some order out of that chaos, eliminated many dishonest schools and incompetent teachers, and vastly improved equipment, curricula, and methods. This should not be forgotten.

But it is easy to see now that the gospel of standardization

was based in part on a tacit, uncriticized, and unwarranted assumption: the assumption, namely, that all men and particularly all children are equal and alike, or nearly equal and nearly alike, not only in their right to Life, Liberty, and the Pursuit of Happiness, as the democratic doctrine declares, but also in kind and degree of intelligence and capacity. In short, we quite overlooked the little matter of Individual Differences, of which, in fact, little or nothing was heard thirty and forty years ago. And we who were making the Standards were, of course, educators, automatically selected in the main on the basis of considerable scholastic or bookish aptitude. Quite naturally, and quite unconsciously, we created our Standards in our own image. We provided for such equipment and curricula and methods as would have been fine for us when we were in school, and prescribed such degrees of attainment as we should triumphantly and joyfully have met if they had been set for us. And all this was excellent and greatly beneficial for that part of the oncoming generation which was like us. But we entirely missed the fact that the great majority of the children in schools, and even a substantial minority of the undergraduates in colleges, were not at all like us, but were endowed with quite other kinds of capacity and often with lesser degrees of capacity of any kind.

For those others — the majority — our generalized uniform Standards were all wrong, in that they gave exclusive sanction and exclusive prestige to tasks which were unsuited to their kinds of capacity or impossible for their degrees of capacity or both. As a consequence, the Standards have caused, and are now causing, untold damage and untellable misery to vast numbers of children in the elementary schools and high schools and even in colleges, thwarting and warping and beating down young lives.

We vaguely imagined that in providing and enforcing sub-

stantially the same kind of instruction for all children we were serving the democratic principle already cited: Life, Liberty, and the Pursuit of Happiness for everybody. But in fact our uniform and exclusively intellectual Standards have deprived a majority of our pupils of the last two of those rights!

In extreme cases children are provoked — by our Standards indirectly — to unsocial rebellion. Senator Copeland has recently emphasized the hideous fact that the present average age of criminals in this country is 23, with the largest age group at 19 and the next largest at 18, and has very gently but plainly brought home a partial responsibility for this situation to the schools and their Uniform Standards.

“The reports of all school systems that have come to my notice,” he writes, “reveal an appallingly large number of academic ‘failures’ in every grade year after year. Authentic testimony indicates that many if not most pre-delinquents are found in these ‘failing’ groups. Are these failures inevitable, or are they due largely to the fact that our curriculum is still so rigid that many of our pupils are confronted with academic tasks which are beyond their abilities, irrelevant to their interests and needs, and which foredoom them to what our inflexible academic standards call ‘failure’?”²

We did not foresee when we made up our beautiful Standards and proceeded to enforce them so firmly that we were about to

² “Education and the Prevention of Crime,” address delivered at the meeting of the Department of Superintendence, National Education Association, Cleveland, Ohio, February 28, 1934; published in *Educational Record*, April, 1934. This address was based in part on the investigation of crime and racketeering conducted by a Senate committee of which Dr. Copeland is chairman. Out of the same investigation has come the Washington Experiment in Character Education, described by Superintendent Frank W. Ballou at the Third Educational Conference, New York City, November 1, 1934; Dr. Ballou’s address has been published in the *Report of the Third Educational Conference*, American Council on Education, 1935.

contribute to a wave of juvenile crime! But even worse, because much more widespread, is the less lurid effect on the vast masses of children who are not driven to crime but only to partial frustration, discouragement, futility, boredom, and various kinds of "escape," into daydreams or into frivolous and unsatisfying distractions outside of school hours.

In sum: the principle of Standards was a thesis which overlooked an equally valid antithesis, the principle of Individual Differences. In the dialectic of educational history that antithetical principle has now come to the fore, and we are forced to construct a synthesis that shall embody the valid aspects of both principles. And since examining has played a major part in the enforcement of Standards, some wiser use of examinations may be vital to the working out of our new synthesis.

Can we indicate briefly what the new synthesis is likely to be, and what kind of examining we shall need in connection with it?

As has already been emphasized, Standards — meaning bookish standards and high bookish standards — are fine for those for whom they are fine; for boys and girls and young men and women possessing a superior degree of scholastic or bookish ability. Our grievous error in this matter of Standards was merely that we conceived of a single standard uniformly applicable to the whole school population.

Plainly, then, what we need is *more* standards; many highly differentiated and carefully graded standards, adapted to as many kinds of capacity and to as many levels of attainment as we can identify. Each of the new differentiated standards would naturally like our present Standards carry its appropriate prescription or indication of subject matter or kind and method of instruction and its own norms of excellence. All should be given equal sanction, and to each should be accorded

its appropriate prestige. Thus, and thus only, shall we succeed in bringing to all children those benefits — first-rate facilities and feasible goals and successful and happy attainment — which our old Uniform Standards sought to bring to all but have actually brought only to one limited group, namely, those who are in some degree bookishly superior.

For the stimulation of achievement and the demonstration of success within each of our new differentiated standards, we shall undoubtedly use appropriately differentiated achievement examinations, both old-type and new-type; though for this delicate business our tests will certainly have to be more carefully constructed for greater reliability than most of our essay-type examinations have had in the past.

But it becomes clear upon very brief reflection that the major rôle of examining in any such set-up will be a new one: not that of determining, at the end of an educational process, whether a common goal has been attained, but rather of discovering, at the beginning of the process and at various points throughout, what different goals should be aimed at.

In short, under a differentiated educational system, adapted to all classifiable kinds and degrees of individual differences, the old major rôle of examinations, the maintenance of standards, will remain, but will become minor. The emergent major rôle of examining is clearly *guidance*; in which process we shall employ both achievement tests and all those other, new kinds of tests — of general intelligence, special aptitude, interest, and personality — for which, as noted above, we have been unable so far to find any very clear use in schools. We can see now that the reason we have been unable to make much use of these other tests is that even when we have their results and recognize those results as valid in the individual case we can seldom, under our present rigid system, do anything about them, so far at least as school programs are concerned.

5. Justifiable and Non-Justifiable Standards-Enforcement

But where does this analysis of the standardizing use of achievement examinations lead us with respect to the appropriateness and rightness of those parts of our present vast examining process in which the maintenance of standards is clearly the major purpose?

It leads us, I think, to justification of some of these items, with some qualifications, and to equally plain condemnation of others.

The clearest case for justification is probably the professional licensing examination, for lawyers, teachers, physicians, and all the way down to midwives and barbers. In such examining we are not concerned primarily with education and the welfare of the individual, but with the protection of society against incompetence and malpractice. Accordingly, such examinations may properly prescribe appropriate subject matter and skills and enforce high standards of attainment; and the standards may well be raised ever higher, as fast as the development of the professions requires and the resources of society permit.

But even here there are two qualifications to be noted:

First, that a high rate of failure in bar, medical, or other professional examinations, such as often prevails, is not a cause for satisfaction but a symptom of gross maladjustment in the system as a whole. This is not to suggest that the candidates who fail should be accepted; it is more likely that an appreciable number who "pass" should be rejected. But it is to suggest that the incompetence of nearly all the failures should have been discovered long before they reached the professional examination and the candidates steered into other fields; not encouraged or permitted to waste their time and money in prolonged preparation for such final defeat. The unsuccessful candidate in a bar examination, for example, has com-

monly had 19 years of general and special schooling. What blank ignorance of our student or bland unconcern with his purposes and welfare is revealed by the fact that through nearly two decades nobody noticed that he lacked the capabilities to make a lawyer, or, having noted this circumstance, concerned himself with it! Such an example may serve to drive home the need for the new emergent function of examining, namely, guidance.

Secondly, I am told by experts in testing that no single examination or brief series of examinations which they have ever devised or expect to succeed in devising attains anything like full reliability for individual cases. If this be so, it would appear that the weight put upon many professional examinations is more than they ought to bear, scientifically or equitably. It does not follow that such examinations ought not to be given, but it does seem to follow that their results should be used, to a larger extent than at present, in connection with other accumulated evidence of the candidates' qualifications and competence. We shall have occasion to return to this point.

What has been said above in justification of the selective, standards-maintaining function of professional licensing examinations may be predicated also, *with the same qualifications*, of examinations for admission, promotion and graduation in professional schools, whether graduate or undergraduate. In these cases also the public interest must be paramount.

On a somewhat different basis a similar justification must be conceded to the endowed colleges and to private secondary and elementary schools. Any such school or college has a clear legal and moral right to choose the clientele it wishes to serve and to offer its facilities to that group only; only, for example, to the bookishly superior. And, having so chosen, any such school or college may properly make use of selective,

standards-maintaining examinations in connection with admission, promotion, and graduation. The process of selective admission especially, in such cases, is in fact a process of educational guidance: guiding into the institution those who may be expected to profit by its facilities, and guiding away, if this be not too sardonic a euphemism for rejection, those who cannot be expected to be successful in the kind of work that is offered.

There is, however, a heavy responsibility upon any such institution to see to it that its selective, standards-maintaining procedures shall have the highest possible reliability for the individual case; because every time those procedures slip serious damage is done to the student victim of the unreliability. Everyone knows that at present the reliability of those procedures is grievously low. The fact that of the students admitted to colleges in any September only about one-half will be graduated four years later is evidence on this point. (Of course, a number drop out for unforeseeable, non-academic reasons — ill health, financial difficulties, and the like; but the residue who are ejected or leave voluntarily because they are misfits is very large.) This situation is clearly due to the fact, that admissions, whether by College Board examinations, or Regents' examinations, or entrance examinations given by the individual college, or certificate, are based on too little evidence, *and* that retention or ejection after admission is also based on too little evidence; in both cases on a very few measurements of low reliability. As a result, even the sacred business of maintaining standards is botched, because of taking too little trouble, because we rely on evidence that no testing expert would accept as valid. And both the guidance function of the admissions procedure, which should be its primary aspect, and educational guidance after admission are almost ignored; are in fact nearly impossible because of inadequate data. A

suggestion for a more adequate method of admissions; *a better maintenance of standards*, serving also as a basis for educational guidance within the college, will be developed in a later section of this paper.

When we come to the schools, of every level, which are supported by public taxation, the legal and moral situation becomes entirely different, and *uniform* standards-enforcing examinations must plainly be condemned. These schools are maintained, as so many orators have boasted, to provide education "for all the children of all the people." And we do not do that by setting up Uniform Standards and standards-enforcing examinations, based upon tasks which are useful and goals which are feasible for only one limited kind and degree of capacity; and, through the agency of such examinations, "flunking," demoting, or retarding all other kinds and degrees. It seems clear that we cannot make further progress in our public schools, elementary and secondary, towards a really democratic education or a really effective education for citizenship until we are prepared to set up a multiplicity of differentiated standards, as suggested above. For really democratic public education must be the exact opposite of the fallaciously democratic *uniform* education we have heretofore sought to impose.

The category of public education includes, of course, the tax-supported state universities and state colleges. These institutions commonly feel themselves to be in a cruel dilemma in the matter of Standards. They are practically all required by law, or by public opinion which would quickly crystallize into law if flouted, to admit all or substantially all graduates of the public high schools of the state, a group which obviously includes varying kinds and degrees of capacity. On the other hand, their administrations and faculties usually hold, with almost religious fervor, that they must uphold Standards —

meaning uniform, high, bookish standards — comparable to those of the great endowed universities.

The practical solution of this *impasse* which some of them have adopted is to receive nearly all comers, as required by law or public opinion, but to effect, shortly after admission, a wholesale elimination, by maintaining in the college courses, through the agency, largely, of examinations and quizzes, Standards which the “non-college material” cannot meet. This is one of the most brutal and indefensible procedures that any type of education has ever sponsored.

And the dilemma which has led to it exists only in the minds of faculties and administrations. By what commandment, other than that of blind imitativeness and conformity, are they required to maintain *for all their students* the kind of Standards that the endowed universities, with their entirely different obligations and responsibilities, have set up?

Certainly a state university should give courses and curricula designed for young people of superior bookish ability, of whom they receive a considerable number, and in those courses and curricula they should maintain Standards second to none; and they must maintain in their professional schools, both graduate and undergraduate, the standards which the respective professions demand; but clearly they should receive into such curricula and schools only those students who have the requisite kind and degree of capacity; and there is no reason why in their huge organizations they should not provide other courses and curricula definitely adapted to other kinds and lower degrees of capacity, in which appropriately differentiated and lowered standards of attainment should be set. They will not be doing a humane or honest job until they either effect such differentiation or else obtain legislative approval for genuinely selective admission and enforce such admission.

So far as I know there is just one state institution which has

looked this problem in the face and made progress towards its solution, namely, the University of Minnesota, in the establishment of its General College, embodying this principle of differentiation of offerings and standards.³

6. Examinations as an Incentive to Study

Our survey of the first uses ordinarily assigned to achievement examinations, namely, the maintenance of standards and selection, has taken us over considerable ground; but this may be proper enough since these related functions are undoubtedly the dominant ones in most current examining. The other alleged uses may be much more briefly reviewed, with one exception.

The third utility of examinations listed earlier in this article was, it may be remembered, that of providing an *incentive to study*.

This point has been conclusively disposed of by Mrs. Eleanor Perry Wood as follows:

"There is little doubt that students are frequently, if not always, stimulated to greater efforts by examinations, but it is a question whether very many receive any genuinely educational motivation from the examination experiences which they have under the present system. On the contrary, we are all aware of the fact that many students are stimulated and provoked to types of conduct and of habits which are not only uneducational, but in some instances morally as well as intellectually negative. Genuine educational motivation is a compound of the individual pupil's abilities and effective interests, and no motivation can long persist which is not fed by consciousness on the part of the student that what he is

³ See J. B. Johnston. *Education for Democracy*. University of Minnesota Press, 1934, especially chapter vi, "New Demands for Differential Treatment of Students in the College of Liberal Arts," and chapter vii, "The Development of Differential College Curricula."

doing is significant, and that he is doing it with visible success and with satisfaction to himself.”⁴

Putting the matter another way: an examination may constitute a minor, temporary incentive of real educational quality in cases where the student's task is congenial to his interests and capacities; it provides an occasion for review and summation and a pleasing opportunity for the demonstration of success to himself and his teachers. But what I have elsewhere called the “policial quiz,” designed to enforce at least periodical industry upon unwilling students, can result only in “cribs” — very few of them technically dishonest, in the form of notes brought in; nearly all of them in the students' heads, where they have been “crammed” the night before. Morally there is supposed to be a vital distinction between these external and internal “cribs,” but so far as effective learning goes there is no substantial difference; in either case the student knows practically nothing about the course a few hours after he leaves the examination room.

7. Examinations as a Method of Instruction

A fourth utility attributed to examinations is that they constitute *a method of instruction*.

An impressive phrase; but in many cases where it is used one finds it difficult to give it specific content. In some instances it seems to refer to the incentive-to-study idea already discussed; particularly to the idea that an impending examination may stimulate students to summarize and organize considerable bodies of material covered over a period of weeks or months. It has been granted above that this may be an incidental value of examinations for students who are well placed in the work they are doing.

⁴ Eleanor Perry Wood. “Examining the Uses of Examinations.” *Harvard Teachers Record*, April, 1933.

In other cases the idea seems to be that through examination papers, corrected and returned, students may be apprised of gaps in their knowledge or errors in their thinking; which is certainly true. But unfortunately the great bulk of essay-type achievement tests are so meager in their sampling that the assistance thus afforded to the student is extremely incomplete and random. A considerable number of other gaps and errors which he might have revealed, if he had had what he would have called the bad luck to be asked other questions, are not brought to his attention. Moreover, when the primary purposes of an examination are those of standards-enforcement and selection, i.e., passing or failing, the psychology of the situation is against much realization of this learning value. The student is likely to regret a mistake not because of the gap revealed in his fabric of knowledge but because of the ten per cent cut from his grade. And since the ten per cent is irremediably gone, why bother much about the gap?

It seems to follow that where it is really desired to use examinations as a method of instruction — where, in other words, this idea is more than a rationalization — the examinations should be given definitely and exclusively for this purpose and should be emphatically dissociated from standards-enforcement; that is to say, they should be corrected but not graded; or, if they are graded, it should be most explicitly understood that the grades are solely for the students' own information as to their measure of attainment and will not be recorded or used in any way in connection with passing, promotion, honors, eligibility, or the like. Any teacher who sets out to use tests in this way is likely to find himself modifying his examination papers considerably with respect to scope and type of sampling, and will encounter also a startling metamorphosis in the attitude of his students towards examinations; he will find that

they actually relish tests of this kind as interesting and helpful occasions.

This has been the experience of scattered teachers in various schools and colleges who have occasionally experimented with this kind of testing; and it has recently been demonstrated on a large scale under the New Plan at the University of Chicago, as reported in Dean Boucher's new book, in which a number of instances are given and are summarized as follows:⁵

"Though the official Board examinations (for promotion to the Senior College and for graduation) are the only required examinations, it was agreed when the New Plan was adopted that reviews, tests, quizzes, and examinations of various types should be given at times and in amounts as needed to attain desired educational results, the need to be determined by faculty and student judgments in the light of experience as each course progressed. Interestingly enough, in not a few courses students have asked that examinations be given more frequently than the instructors thought necessary to acquaint both students and instructors adequately with the rate and degree of progress being made by the students. The tendency is to accede to student requests in such instances. In more than one instance at the end of the Autumn and Winter quarters, after several instructional tests have been given during the quarter upon the conclusion of logical units of work, it has been left to student vote to determine whether a final examination on the entire quarter's work should be given. In every such instance the students have asked for the examination, though they knew that the result would have no officially recorded effect upon their attainment of the junior-college certificate. They did know, however, that the examination papers

⁵ Chauncey Samuel Boucher. *The Chicago College Plan*, pp. 115, 116, (1935). Reprinted by permission of the University of Chicago Press.

would be carefully corrected and returned, and would thus serve as an excellent instructional aid in their endeavor to master as much of the field as possible in preparation for the official Board examinations."

Here we have the real thing in the use of examinations as a method of instruction, with the necessary complete dissociation from the standards-enforcement function; which function is taken care of separately at Chicago by the official Board examinations.

8. Do Examinations Conduce to Improvement of Teaching?

Next we may consider the value of examinations as a means of stimulating or enforcing *improvement of teaching*.

This idea also seems to have diverse meanings in different mouths. Where it is invoked in connection with such procedures as those of the College Entrance Examination Board and the New York Regents it seems to mean that through the agency of centralized examinations teachers have been led to select materials of instruction which are deemed desirable, to cover a desiderated amount of such materials, and to be thorough and meticulous in their work, or at least in such aspects of it as will contribute to a good showing in the examinations. But the improvement of teaching in this sense is obviously synonymous with the enforcement of standards; which objective or use of examinations has already been discussed with some fullness, except for one point, which is highly relevant under this heading.

That point is the contention of a considerable number of teachers, especially those of the Progressive persuasion, that standards-enforcing examinations, far from contributing to the improvement of teaching, lead to its deterioration; that they force teachers into patterns of routine, compel them to ignore individual differences in their students, kill their spon-

- taneity and initiative, and above all cause them to concentrate on factual material, mere information (to be used in passing the examinations), to the neglect of the vital goals of education, such as growth and power and understandings and appreciations and socialized attitudes.

In my own view there has been historically considerable basis for this objection. Certainly a great many teachers who have taught under the College Board system or the New York Regents' system are wont to proclaim it loudly. From this point of view the item *Improvement of Teaching* should be stricken from the list of the uses of examinations, and we should set up a new list of *disutilities* of examinations, headed, perhaps, with the item Stultification of Teaching.

Nevertheless it is perfectly clear that we are not going to abandon examinations. They are necessary instruments of too many important educational and social purposes. Accordingly the best we can do is to minimize as far as possible this aspect of disutility. It will be noted that this aspect inheres chiefly in the standards-enforcing and selective uses of examinations. It does not appear at all, for example, in their genuine, separate use as a method of instruction, nor yet in the remaining major use of examinations, already suggested and to be discussed below, namely, guidance. Moreover, it will tend to disappear, even in connection with standards-enforcing examinations:

1. As such examinations become more comprehensive, covering larger areas, with more freedom allowed. (The old charges are seldom if ever brought against the College Board's new "English Comprehensive.")

2. If and when we set up a multiplicity of differentiated standards in place of our present Uniform Standards.

3. As we come to use the results even of standards-enforcing examinations less as isolated measures and more in connection

with accumulated records of the students' capacities and previous achievements.

9. The Appraisal of Teachers and Departments

School and college administrators have long played with the idea that a study of examination results, of the distribution of the grades given by different instructors, might be made to yield some evidence of the comparative teaching effectiveness of the members of their staffs. In the past this idea has usually been abandoned after brief consideration because of the notorious lack of comparability of the old essay-type examinations; but with the advent of the objective and standardized new-type tests it recurs very temptingly.

It is tempting for perfectly legitimate reasons. In the elementary and secondary schools it has been recognized all along that teaching effectiveness should be the principal criterion in the selection, retention, and promotion of teachers; hence, a pressing administrative need for measures of pedagogical success much more definite and objective than any we have ever had. Even college and university presidents have admitted the importance of this criterion, and have avowed that they would make use of it more than they do if it could be as definitely appraised as administrative effectiveness or activity in research. And beyond this all experiments looking to the improvement of teaching processes must be somewhat inconclusive until we can have reliable, comparable measures of results.

It does seem possible that in time methods of realizing this highly desirable theoretical utility of comparable tests may be worked out. But it is clear also that for the present any procedures in this direction should be entered upon with caution and watchful skepticism and should be regarded as experimental. No simple comparison of the percentile rankings of the same students in different courses, for example, or of the

average or median scores of different classes in the same course, will yield safe interpretations, no matter how objective and reliable the test or tests may be. We must await the development of controls and techniques as yet only partially envisaged.

Some of the items which will enter into an eventual technique can, of course, be foreseen. One such item will undoubtedly be the checking of achievement test results against intelligence test scores; this procedure is already generally followed, and obviously increases the reliability of measurement or indication of probable effectiveness of instruction. As we come to have a larger number of special aptitude tests, a check of achievement test results against specific aptitude scores, as well as against general intelligence scores, should further increase the probability of reliable indication. For example, achievement tests in music could now be checked, not only against general intelligence, but against the Seashore Measures of Musical Talent; and achievement tests in physics against the arithmetical and analogies subdivisions of the American Council on Education Scholastic Aptitude Test, since scores in these subtests have been found to correlate well with success in physical science.

But it seems certain that still further controls, including careful definition of necessary limiting conditions, will have to be worked out before administrators will have that dependable yardstick which they so greatly need in this matter.

Even now, however, individual teachers may profitably use examination results, especially on standardized tests, for *self-appraisal*, including the discovery of some of their own shortcomings, such as failure to include or emphasize important topics or to explain terms and principles with clarity. Such individuals can be trusted to give themselves the benefit of the grave technical uncertainties involved, whereas administrators, even department heads, dealing with others than them-

selves, might often be tempted away from the necessary caution and come to treat as definite evidence results which would have at most some uncertain probability.

This seems to me about all that can be safely recommended at present under this head, except, of course, in the case of trained investigators thoroughly familiar with testing procedures and the statistical interpretation of test results. But, as already noted, we may reasonably hope for further help at this point as the science and art of testing progresses.

10. The Accrediting of Schools and Colleges

Now that we are all dissatisfied with the old quantitative standards for accrediting schools and colleges which were set up twenty-five years ago, we have naturally thought of the possibility of using comparable standardized tests to check up on whole institutions, both secondary and higher. To test the product rather than prescribe the process seems a good idea; and this idea has already been invoked in studies conducted at Indiana University and at the State University of Iowa and in the elaborate investigation of accrediting procedures sponsored by the North Central Association of Colleges and Secondary Schools, which has resulted in that Association's quite revolutionary new qualitative criteria. In the course of that study, "an extended testing program was engaged in, and institutions were rated according to each test and according to a composite test score."⁶ Some idea of rating institutions on the basis of achievement test results was implicit also, of course, in the procedures of the Pennsylvania Study.

But this use of examinations is clearly subject to caveats similar to those already presented in connection with their use for the appraisal of departments or of individual teachers.

⁶ Dr. George F. Zook, "Accreditation of Secondary Schools in the Light of the North Central Association Report." *Educational Record*, January, 1935.

This is evidently the conclusion reached by North Central Association's commission, and the whole argument has been so admirably presented by Dr. Zook in a report of that commission's work that I shall take the liberty of quoting him at length:⁷

At this point may I digress to comment upon a line of discussion which has been interpreted by many people as removing all reason for accrediting colleges and schools? I refer to the development of the testing and personnel movements, which it is said substitute the accrediting of students as individuals for the accrediting of institutions. Such an argument seems very intriguing at first, but I am convinced that conclusions much too far-reaching in character have been drawn from the studies which have been made.

Recent years have seen a tremendous advance in devising scientific measures for testing the native ability and the achievement of individual students. All of us rejoice at the progress which has been made in this field. We believe that the testing movement will gradually enable us to solve many of our most difficult problems in counseling, guidance, and placement. If I err in my attitude toward this new device of education it is the same error which many other ardent friends of the movement commit, namely, that of placing far greater reliance on the results of tests than dispassionate judgment probably justifies. I mention this because I do not wish my position relative to the value of the testing movement to be misunderstood.

The researches of Professor H. H. Remmers in Indiana and the group at the State University of Iowa, for example, are in point. In both of these states through an extensive testing program of students enrolled in the secondary schools it was demonstrated, first, that there are marked differences between the achievement of students in the various schools, and that the present methods of accrediting schools did not by any means separate out the schools whose pupils on the average achieved the most from those

⁷ Dr. George F. Zook, address on "Accrediting Schools and Colleges," delivered at the Second Joint Educational Conference, New York City, November 2, 1933, and published in the report of the Conference, *Educational Measurement and Guidance*, American Council on Education, Washington, D.C., 1933.

which had lower averages. Furthermore, even the schools with the lowest averages had considerable percentages of students in the upper brackets, while the best always contain many who do not show up well. . . .

It should be admitted at once, as I have endeavored to do elsewhere in this paper, that the methods used by the accrediting agencies in evaluating the effectiveness of a school or college have been crude and unscientific, but it does not follow, simply because there is a wide difference between two schools as to the levels which their students reach respectively on achievement tests, that one is a better or a more effective school than the other. Any two schools will vary greatly in types of students which they enroll just as any single school will contain students with a rather wide range of ability. Hence to assume that only those schools or colleges whose students make the highest average on tests are the ones which should receive recognition is to ignore altogether both the level of native ability and of preparation which students possess at the time of entrance. A school or a college may enroll students of fairly high ability and good preparation and yet do relatively a poor piece of work. On the other hand, a school or college enrolling students whose level of native ability and preparation may be fairly low and which adapts its resources vigorously to a program suited to the needs of its students may be in its realm as effective if not more effective than the first institution. Until, therefore, we relate the results of tests for native ability to the results of achievement tests from the same students, until we compare the achievement of students at the end of two, three, or four years with what they started out with, and until we take into account the differences in objectives which the respective schools and colleges may be expected to have, we have no right to assume that a high level of achievement test scores is any better gauge of the effectiveness of the school or college than the crude methods now being followed by the accrediting agencies as a means of identifying the effectiveness or the quality of the institution. So far as I know the conclusions which have been reached relative to the substitution of the accrediting of individuals for the accrediting of institutions have not been based on considerations of this kind.

When that stage of development is reached in the testing pro-

gram the results of tests may become a significant means of identifying the quality and the effectiveness of the educational process at a given school or college in comparison with what is found through the same process at another school or college. To whatever extent, therefore, tests can be used in the process of accrediting schools and colleges, I am sure that they will be welcomed as one of the few objective and scientific means of evaluating an institution that we now have at our disposal.

On the other hand, not even the most ardent friend of the testing movement would claim that the results of tests measure all that we expect our youth to get out of school or college. An English schoolmaster recently said that he believed that the education which a school boy receives out of class was more significant than what he secures in the classroom. Whether we agree with him or not we must agree that tests do not measure this aspect of the educational process very well, if at all.

In the same way the results of tests do not measure all that we expect a school or a college to be. Fond fathers and mothers, for example, want to know whether the social life of the school or college is healthful; the people of a community want to know whether their schools have adequate equipment to do the things they attempt to do; the constituency of a school or college wants to know whether the finances are being handled properly; in short, society is interested in the total pattern of an educational institution, including that part of the educational process which can be measured objectively, that part which is imponderable, and even that administrative and physical machinery and equipment commonly regarded as necessary or desirable for an effective school or college.

I trust, therefore, that you will join me in regarding the accrediting of students through the testing and personnel movements not as a substitute for the accrediting of schools and colleges but as a supplement to it. Both are extremely valuable and important social devices operating for essentially different though supplementary purposes. While the accrediting of individuals may modify the methods of accrediting institutions and even be used as an important means to that end, it will never be more than one of those means. Society is anxious for the new ways of identifying the native ability, the special interests, and the

achievement of individuals, but it is equally interested in ways and means of identifying the total effectiveness and the adequacy of the institutions which it establishes for these and other purposes incidental thereto.

Later in the same address Dr. Zook presents what seems to be the nub of the matter in the following sentences:

The testing movement may prove very helpful, but I wish now to hazard the guess that the accrediting associations should not themselves engage in a comprehensive testing program in institutions which are applying for accrediting. They may, however, properly expect an institution to show that it has an active testing program and to make the results available.

In other words, self-appraisal by an individual institution, like an individual teacher, on the basis of test results, is probably feasible and worth while; but for some time to come at least testing should play only an incidental and auxiliary part in general accrediting procedures.

11. The Emergent Major Function of Examinations: Guidance

There remain two other utilities which have been claimed for examinations: that they furnish data for educational guidance, and that they accumulate materials for research. But since everyone would admit, I judge, that the latter of these functions must usually be incidental and secondary, we need discuss here only the former: the use of examinations in connection with *educational guidance*, which I have already characterized as their emergent major function.

The reason why guidance is the emergent major function of examining is, of course, that it is the emergent major function of the entire educational process. At long last we are beginning to understand that we must learn our pupils before we can teach them; that unless we know quite a lot about a student's abilities, aptitudes, interests, and personality, and

his previous attainments, and his background and circumstances, we cannot, except by blind chance, select the teaching materials, i.e., "courses," or methods or disciplines or goals which will be really serviceable to him; that in the absence of such knowledge we are very likely to impose upon him materials, methods, disciplines, and goals that will be thwarting and stultifying; that in the light of such knowledge, on the other hand, we can, or under a sufficiently flexible system could, plan for him an educational program into which he would throw himself with zest, and in which he could work on his own power and, almost independently of us, attain self-education, which, as we so often announce, is the only real education.

All of us have now and then the experience of steering some student who has been failing and miserable in a program for which he was not well adapted into some other program which really fits his needs, and seeing him blossom out into happy success. One problem case solved! But we realize now that we ought to be doing this, not now and then, but all the time with all our students at every level. In short we are coming to appreciate the literal correctness of Professor Morrison's startling dictum "to the effect that teachers should spend half their time studying their pupils as growing individuals, and the rest of their time doing what that study indicates is desirable and necessary."⁸

⁸ Cited from Dr. Ben D. Wood's "The Major Strategy of Guidance" (*Educational Record*, October, 1934). Readers interested in educational guidance and the relation of examinations to guidance are referred to that article and to two other important articles by Professor Wood: "Basic Considerations in Educational Testing" (read at the National Education Association meeting in Minneapolis, February 28, 1933, and published by the Committee on Educational Testing, May, 1933); and "The Ultimate Basis for Satisfactory College-High School Relations" (*Bulletin of the American Association of Collegiate Registrars*, July, 1934). Reprints of these three articles may be obtained from the Cooperative Test Service, 500 West 116th Street, New York City.

We are beginning to see also that this doctrine has two major corollaries:

First, *a very great diversification of our courses and methods and disciplines and goals*; the setting up of many differentiated standards, adapted to all classifiable kinds and degrees of Individual Differences. The theoretical ideal is an individual curriculum built for each individual boy and girl. This ideal we shall never fully attain for practical reasons of administration and expense. But there is no reason why it should not be nearly approached through an adequate increase in the variety of courses and methods of instruction in our existing schools and in new kinds of schools which should be created, *plus* a free and flexible utilization of this variety of courses and methods in the interest of each individual. The increased expense involved could be met from a fraction of the current expenditures of society upon that considerable proportion of crime, unemployment, and destitution which is directly traceable to the malfeasances of our present rigid educational system.

Second, *the general introduction into the schools throughout the Nation of methods now available for the study of individuals and individual guidance*.

In practice this second corollary will have to be developed first, because it is the easier and cheaper to develop, and because only the actual study of many thousands of individual students in many hundreds of schools and unavailing efforts to provide for the ascertained needs of those individuals within our present rigidities will bring home even to educators, not to mention boards of education and the tax-paying public, the desperate need for more diversified facilities.

It has to be conceded, of course, that the methods referred to as being now available for the study of individuals, with a view to educational and vocational guidance, leave much to

be desired, are capable of vast improvement. But the fact is that this aspect of the matter has been unduly stressed, and has been stressed largely as a defense mechanism, so that we might feel satisfied to postpone the effort (and expense) of introducing such methods into our own schools.

The other side of the truth, to which so many of us try to close our eyes, is that the methods already known of learning about boys and girls, ascertaining their abilities and aptitudes and interests and qualities and backgrounds and circumstances, and assembling such information for intelligent and highly useful use, are sufficient to enable us teachers and administrators to avoid at least many of our habitual mistakes and educational crimes, if only we will make use of them! Does a physician wait until he has a specific for a disease or an absolutely perfect method of diagnosis and therapy before he will undertake a case?

Which brings us back to our central topic of examinations and their use in educational guidance. Guidance has use not only for achievement tests, the only kind we use for all other purposes of examinations, but also, as previously noted, for the other, new kinds of tests which the psychologists and testing experts have developed: intelligence tests, special aptitude tests, vocational and avocational interest tests, and personality ratings.

In connection with achievement testing for guidance purposes three points may be made:

1. For the guidance purpose, even more vitally than for other purposes, a high degree of comparability of test results is important. We need to know, not only how well Johnny has done in meeting or excelling the somewhat vague minimum or "passing" standard of some particular teacher or even of some central examining board, but chiefly how he ranks, in his mastery of various materials of learning, in comparison with large groups of other students with similar and dis-

similar backgrounds from similar and dissimilar schools. This is the essential point which achievement tests can yield for purposes of guidance; that is, as a basis of judgment of ability and aptitude and as to the kind of school that should be attended, the materials and methods of instruction that will be profitable, and ultimately the occupation which may wisely be aimed at. It follows that, for guidance, the new-type, standardized, objective achievement test, because of the vastly greater comparability of its results, is much better than the old essay-type examination. Hence, the great importance for educational guidance of such series of comparable tests as are now provided by the Cooperative Test Service,⁹ and of such broad testing programs as those sponsored by the Educational Records Bureau and the state-wide testing programs developed in Minnesota, Wisconsin, Ohio, Iowa, Colorado, Alabama, North Carolina, Kentucky, and other states.

2. Achievement testing for guidance serves its purpose most clearly and efficiently when it is dissociated from other purposes, especially from standards-enforcement and invidious selection; just as we have already seen that a real and effective use of examinations as a method of instruction must be dissociated from standards-enforcement. The obvious reason lies in the inescapable difference between a student's willing cooperativeness towards an examination the results of which he is assured will be used *for* him, to help him solve his own problems, and his instinctive effort to "beat" any examination whose results are going to be used *against* him, in matters of admission, promotion, or graduation. We may compare the ready willingness of anyone to have his blood pressure taken by his family physician as against the attitude with which the

⁹ See H. E. Hawkes, "The Cooperative Test Service," in *Educational Record*, January, 1931, or Max McConn, "The Cooperative Test Service," in *Journal of Higher Education*, May, 1931.

most upright among us would be likely to confront the blood-pressure lie-detector.

3. For the guidance purpose we need a large number of tests of each individual, for the reason already cited, that no single test, and no group of tests to be administered on one crucial occasion, can give us a valid diagnosis of an individual's capabilities or even of his achievements. For such diagnosis, and hence for educational guidance, we need, as I have written elsewhere, "*a series of comparable tests over a period of years*, through which the stresses and accidents of particular occasions may cancel, and by means of which we may follow both persistent levels and significant changes of capacity, interest, and attainment."¹⁰

The foregoing paragraphs contemplate the collection of a large amount of data for each student, including the results of many tests of many kinds over a period of years. With such data must be collated all the definite biographical material the student's adviser can obtain to illuminate his background and circumstances. Obviously we have need of some instrument to assemble this mass of information in clear and significant fashion; and fortunately such an instrument is at hand in the Cumulative Record Card, whether in the original form published by the American Council on Education, or in the adaptation used by the Educational Records Bureau, or in other variants developed in many schools and colleges throughout the country.

"These blanks or folders," if I may quote an earlier description, "provide an instrument for organizing and presenting, compactly and in part graphically, on a time projection, all the significant facts, both scholastic and personal, in regard to a student's career. . . . These significant facts include not

¹⁰ "Educational Guidance Is Now Possible." *Educational Record*, October, 1933.

only school marks and the results of objective tests but also such items as health, physical and mental, family background, financial situation, study conditions and programs, extra-curricular activities, summer experiences, vocational experiences, unusual accomplishments, reported interests, educational and vocational plans, and the like, all carried forward from year to year to exhibit both permanent and changing conditions and tendencies."¹¹

Every teacher or administrator or personnel officer who has used such cumulative records knows that in many cases they are almost self-interpreting; that the multifarious data they contain, compactly and sequentially organized and in part graphed, often fall at once into a clear picture, each part reinforcing every other part, and the whole affording decisive answers to such questions as to what courses the student should take and avoid, what kind of discipline he needs, what further schooling he should seek, and even what kinds of career he should and should not endeavor to follow — answers decisive and convincing not only to the teacher or school officer but also to the student himself and his parents. In other cases, of course, the picture is not so clear; one part of the data may seem to contradict other parts; but even such discrepancies usually afford clues for further investigation or experiment, with more hope of a final solution than any other procedure holds out.

Many of us hope that the general introduction into our

¹¹ "The Cooperative Test Service." *Journal of Higher Education*, May, 1931. Full descriptions of these forms, with samples, have been published in the *Supplement to Educational Record* for July, 1928; in the *Program for a Study of the Relations of Secondary and Higher Education* (the "Pennsylvania Study"), issued by the Carnegie Foundation for the Advancement of Teaching, 1928; and in an article on the Pennsylvania Study by the present writer in the *Bulletin of the American Association of Collegiate Registrars*, vol. v, no. 2 (1929). See also a valuable discussion in Dr. Ben D. Wood's "The Major Strategy of Guidance." *Educational Record*, October, 1934.

schools at every level of examinations given and used specifically for guidance, and a parallel general introduction of cumulative records to assemble the results of such examining with all other relevant data, will eventually solve the problem of admission to college and also the problem of guidance in college and with respect to professions, reducing to a minimum failures in college after admission and failures in professional schools and professional licensing examinations.

And we believe also that the increasing use of examinations for guidance and of cumulative records will at last reveal inescapably to all that pressing need, which only a few now perceive, for what I have called the differentiation of standards; involving the provision of many new courses and of new kinds of schools, and greatly increased flexibility in drawing upon the offerings of every kind of school in the interest of individual needs, in order that we may ultimately provide suitable instruction or training, with feasible goals, for every kind and degree of capacity.

12. Summary

I set out in this article to survey the actual uses of examinations in our current educational practice, with the hope of arriving at some clear doctrine as to the right use of this instrumentality. What conclusions may be drawn from the facts and considerations presented?

It would seem to me it has appeared:

1. That the great bulk of our present huge examining activity is achievement testing, using mainly the old essay-type examination, and devoted to the combined purposes of standards-enforcement and selection.

2. That the standards thus enforced are excellent and beneficial for one particular type of boy and girl and young man and young woman, namely, those of superior bookish ability.

3. That those same standards, however, because of their uniformity and rigidity, are thwarting and damaging to all other kinds and degrees of capacity.

4. That we need, therefore, to serve these other kinds and degrees, many more diversified standards, involving new kinds of courses and schools and methods and disciplines and goals.

5. That for the demonstration and maintenance of these new diversified standards we shall need a parallel diversification of achievement tests, of improved reliability.

6. That with the development of such diversification, however, the major use of examinations will come to be, not the enforcement of standards, but guidance.

7. That some parts of our present standards-enforcement examining are justifiable, including examinations for professional licensure and for admission, promotion, and graduation in professional schools, endowed colleges, and private secondary and elementary schools.

8. That even in these cases our present standards-enforcing examinations need support and correction from cumulative records of many previous tests and other data in regard to the students' careers.

9. That the enforcement of *uniform* standards in public schools at any level must be condemned; which is merely a corollary of (3) and (4).

10. That the alleged utility of examinations as an incentive to study is largely an illusion and rationalization; at best a minor, incidental value realized only in cases where the goal set is thoroughly appropriate to the students' capacities.

11. That the utility claimed for examinations as a method of instruction can be genuinely realized only when examinations are used specifically for that end and divorced from standards-enforcement.

12. That the utility of examinations for the improvement

of teaching is either synonymous with the enforcement of standards or highly dubious; contradicted, in fact, by claims that examinations tend to stultify teaching.

13. That such stultifying effects will, however, tend to disappear as we come to use more comprehensive tests, more diversified standards, and cumulative records.

14. That we have not yet developed the necessary technique for the safe general use of examination results for the appraisal of teachers or departments; but that individual teachers may profitably use such results for self-appraisal and self-diagnosis.

15. That, similarly, testing procedures cannot at present, if ever, be relied upon as a sole basis for the accrediting of schools and colleges, but may constitute a valuable item in accrediting, as contemplated in the new North Central Association criteria.

16. That the emergent major utility of examinations is educational guidance.

17. That the developing doctrine of guidance demands, first, diversification of standards and courses and schools, and, second, the general introduction of the methods now available for the study of individuals.

18. That in examining for guidance we need all known kinds of tests, many of them, preferably comparable tests, and preferably tests used explicitly for this purpose.

19. That cumulative records are essential for the summation and interpretation of personnel data, including test results and other relevant information.

20. That the general use of examinations for guidance and of cumulative records will ultimately go far towards solving the problems of admission to college, professional and other vocational guidance, and the needed diversification of standards.

BIBLIOGRAPHY

Richard D. Allen: Inor Group Guidance Series.

Vol. I: *Common Problems in Group Guidance*, 1933.

Vol. II: *Case-Conference Problems in Group Guidance*, 1933.

Vol. III: *Self-Measurement Projects in Group Guidance*, 1934.

Vol. IV: *Organization and Supervision of Guidance in Public Education*, 1934.

New York: Inor Publishing Company (30 Irving Place).

Richard D. Allen: "The Program of Measurement in the Secondary Schools of Providence." *Junior-Senior High School Clearing House*, 8:326-29; February, 1934.

Florence M. Baker: *The Teaching of French*. Boston: Houghton Mifflin Co., 1931.

Loren Bane: Analysis of Every-Pupil Test in United States History of the 1932 Iowa Academic Contest. Unpublished M. A. thesis, State University of Iowa, August, 1932.

Francis F. Bradshaw: "The Scope and Aim of a Personnel Program." *Educational Record*, 17:120-28; January, 1936.

John M. Brewer: *Education as Guidance*. New York: The Macmillan Co., 1933.

C. S. Boucher: *The Chicago College Plan*. Chicago: University of Chicago Press, 1935.

Oscar K. Buros: "Educational, Psychological, and Personality Tests of 1933 and 1934." *Studies in Education*, No. 7; Rutgers University Bulletin, Volume XI, No. 11, May, 1935.

Mary C. Champneys: *An English Bibliography of Examinations*. London: Macmillan and Co., Ltd., 1934.

W. W. Charters: "Education and Research at a Mechanics Institute: A Character Development Study." *Personnel Journal*, 12:119-23; August, 1933.

Frederic D. Cheydeur: "Placement and Attainment Examinations in Foreign Languages." *Educational Record*, 15:176-91; April, 1934.

Frederic D. Cheydeur: "Attainment Examinations in Foreign Languages at the University of Wisconsin." *French Review*, 6:190-214; February, 1933, and 6:282-300; March, 1933.

BIBLIOGRAPHY

- L. E. Cole: "Latin as a Preparation for French and for Spanish." *School and Society*, 19:618-22; May 24, 1924.
- Algernon Coleman: *Experiments and Studies in Modern Language Teaching*. Chicago: University of Chicago Press, 1934.
- Algernon Coleman: *Teaching of Modern Foreign Languages in the United States*, Publications of the American and Canadian Committees on Modern Languages, Volume XII. New York: The Macmillan Co., 1929.
- College Admission and Guidance*, Report of an Educational Conference, New York City, November 3, 1932. Reprinted from *Educational Record* for January, 1933.
- W. W. Cook: "The Measurement of General Spelling Ability Involving Controlled Comparisons Between Techniques." University of Iowa, *Studies in Education*, Vol. VI, No. 6, 1932.
- Royal S. Copeland: "Education and the Prevention of Crime." *Educational Record*, 15:123-37; April, 1934.
- Alice Corell and others: *Tentative Syllabus in Modern Foreign Languages*. Albany, New York: University of the State of New York, 1931.
- A. B. Crawford: "Aptitude Testing in Personnel Procedure." *Bulletin of the American Association of Collegiate Registrars*, pp. 293-309, July, 1934.
- A. B. Crawford: "Some Criticisms of Current Practice in Educational Measurements." *Harvard Teachers Record*, 3:67-81; April, 1933.
- A. B. Crawford: "Forecasting Freshman Achievement." *School and Society*, 31:125-32; January 25, 1930.
- A. B. Crawford and Paul S. Burnham: "Entrance Examinations and College Achievement." *School and Society*, 36:344-52, 378-84; September 10 and September 17, 1932.
- Edgar Dale: Familiarity of 8000 Common Words to Pupils in the Fourth, Sixth, and Eighth Grades. Bureau of Educational Research, Ohio State University. Unpublished.
- S. B. Davis and E. E. Hicks: *Historical Content and Background of Caesar's Gallic War*. Pittsburgh, Pennsylvania: University of Pittsburgh.
- Definition of the Requirements for 1934, with the Examinations for 1933. Office of the Board of Secondary Education, Milton, Massachusetts.

BIBLIOGRAPHY

- Educational Measurement and Guidance*, Report of the Second Educational Conference, New York City, November 2-3, 1933. American Council on Education, 744 Jackson Place, Washington, D.C. (Reprinted in part from the *Educational Record*, January, 1934.)
- Educational Records Bureau: "Bibliography of Reading Tests." Educational Records Bureau, 437 West 59th Street, New York City, January 15, 1935.
- Educational Records Bulletin, No. 11, *Achievement Testing in Public and Independent Secondary Schools*, 1933. Educational Records Bureau, 437 West 59th Street, New York City.
- Educational Records Bulletin, No. 12, *1933 Fall Testing Program in Independent Schools*, Educational Records Bureau, 437 West 59th Street, New York City.
- Educational Records Bulletin, No. 13, *1934 Achievement Test Program in Independent Schools*, Educational Records Bureau, 437 West 59th Street, New York City.
- Educational Records Bulletin, No. 14, *1934 Fall Testing Program in Independent Schools*. Educational Records Bureau, 437 West 59th Street, New York City.
- Educational Records Bulletin, No. 15, *1935 Achievement Test Program in Independent Schools* (Including Statements from Schools on the Uses Made of Comparable Tests and Cumulative Records). Educational Records Bureau, 437 West 59th Street, New York City.
- Educational Tests and Their Uses. Review of Educational Research*, Vol. III, No. 1, February, 1933. (Prepared by the Committee on Educational Tests and Their Uses: Ben D. Wood, W. J. Osburn, G. M. Ruch, M. R. Trabue, Grace A. Kramer, and John L. Stenquist, Chairman; with the assistance of E. F. Lindquist and H. R. Anderson.)
- Hiram J. Eininger: Pupil Information Bearing on Important Topics in American History, 1789-1798. Unpublished M. A. thesis, State University of Iowa, August, 1933.
- H. E. Ford: *Modern Language Instruction in Canada*, Part II. Publications of the American and Canadian Committees on Modern Languages, Vol. VI. Toronto: University of Toronto Press, 1928.
- Fred P. Frutchey: "Measuring the Ability to Apply Chemical

BIBLIOGRAPHY

- Principles." *Educational Research Bulletin*, Ohio State University, 12:255-60; December 13, 1933.
- Arthur I. Gates: *The Improvement of Reading*. New York: The Macmillan Co., 1927.
- O. T. Gooden: "Testing in the College." *Journal of Higher Education*, 7:191-95; April, 1936.
- Catherine M. Haage: Tests of Functional Latin for Secondary School Use. Doctoral thesis, University of Pennsylvania, Philadelphia, 1932.
- Peter Hagboldt: *Building the German Vocabulary*. Chicago: University of Chicago Press, 1930.
- Peter Hagboldt and F. W. Kaufman: *Lesebuch für Anfänger*. Chicago: University of Chicago Press, 1930.
- H. E. Hawkes: "The Cooperative Test Service." *Educational Record*, 12:30-8; January, 1931.
- H. E. Hawkes: "Report on the Cooperative Test Service." *Educational Record*, 14:391-97; July, 1933.
- H. E. Hawkes: "Report on the Cooperative Test Service." *Educational Record*, 15:359-67; July, 1934.
- B. C. Hendricks, R. W. Tyler, and F. P. Frutchey: "Testing Ability to Apply Chemical Principles." *Journal of Chemical Education*, 11:611-13; November, 1934.
- V. A. C. Henmon: "Recent Developments in the Study of Modern Foreign Language Problems." *Modern Language Journal*, 19:187-201; December, 1934.
- V. A. C. Henmon: *Achievement Tests in the Modern Foreign Languages*. Publications of the American and Canadian Committees on Modern Languages, Vol. V. New York: The Macmillan Co., 1929.
- L. J. Henry: "Comparison of the Difficulty and Validity of Achievement Test Items." *Journal of Educational Psychology*, 25:537-41; October, 1934.
- Gertrude Hildreth: *A Bibliography of Mental Tests and Rating Scales*. New York: Psychological Corporation (522 Fifth Avenue), 1933.
- Karl J. Holzinger and Frances Swineford, compilers: "Selected References on Statistics and the Theory of Test Construction." *School Review*, 41:462-66, June, 1933; 42:459-65, June, 1934; 43:462-67, June, 1935.

BIBLIOGRAPHY

- J. B. Johnston: "Advising College Students." *Journal of Higher Education*, 13:15-20; June, 1930.
- J. B. Johnston: *Education for Democracy*. Minneapolis: University of Minnesota Press, 1934.
- J. B. Johnston, chairman: "The 1932 College Sophomore Testing Program." *Educational Record*, 13:290-343; October, 1932.
- J. B. Johnston, chairman: "The 1933 College Sophomore Testing Program." *Educational Record*, 14:522-71; October, 1933.
- J. B. Johnston, chairman: "The 1934 College Sophomore Testing Program." *Educational Record*, 15:471-516; October, 1934.
- J. B. Johnston, chairman: "The 1935 College Sophomore Testing Program." *Educational Record*, 16:444-81; October, 1935.
- E. S. Jones: *Comprehensive Examinations in American Colleges*. An investigation for the Association of American Colleges. New York: The Macmillan Co., 1933.
- C. H. Judd and G. T. Buswell: *Silent Reading, A Study of the Various Types*. Chicago: University of Chicago Press, 1922.
- Truman L. Kelley and A. C. Krey: *Tests and Measurements in the Social Sciences*. Report of the Commission on the Social Studies of the American Historical Association, Part IV. New York: Charles Scribner's Sons, 1934.
- Mary H. King: Pupil Comprehension of Place Location Data in Junior High School American History. Unpublished M. A. thesis, State University of Iowa, July, 1935.
- T. J. Kirby: "Latin as a Preparation for French." *School and Society*, 18:563-69, November 10, 1923.
- A. C. Krey: "The Effect of Measurement on Instruction." *Journal of Educational Research*, 28:498-501; March, 1935.
- W. S. Learned: "Testing for Values in Education." *Bulletin of the Association of American Colleges*, Vol. XX, No. 1, March, 1934.
- W. S. Learned: (Final Report on the Pennsylvania Study) *Bulletin* 28 of the Carnegie Foundation for the Advancement of Teaching, 522 Fifth Avenue, New York City, 1936.
- J. Murray Lee and Percival M. Symonds: "New Type or Objective Tests: A Summary of Recent Investigations (October, 1931, to October, 1933)." *Journal of Educational Psychology*, 25:161-84; March, 1934.
- W. M. Lewis: "Credit Hunting versus Education." *Educational Record*, 13:38-49; January, 1932.

BIBLIOGRAPHY

- Louis H. Limper: "Student Knowledge of Some French-English Cognates." *French Review*, 6:37-49; November, 1932.
- Edward A. Lincoln and L. L. Workman: *Testing the Uses of Test Results*. New York: The Macmillan Co., 1935.
- E. F. Lindquist: "Objective Achievement Test Construction (bibliography)." *Review of Educational Research*, 5:469-83; December, 1935.
- E. F. Lindquist: "The Form of the American History Examination of the Cooperative Test Service." *Educational Record*, 12:459-75; October, 1931.
- E. F. Lindquist and H. R. Anderson: "Achievement Tests in the Social Studies." *Educational Record*, 14:198-256; April, 1933.
- E. F. Lindquist and W. W. Cook: "Experimental Techniques in Test Evaluation." *Journal of Experimental Education*, 1:163-85; March, 1933.
- John A. Long: "Improved Overlapping Methods for Determining Validities of Test Items." *Journal of Experimental Education*, 2:264-68; March, 1934.
- Max McConn: "Educational Guidance Is Now Possible." *Educational Record*, 14:475-99; October, 1933.
- Max McConn: "Educational Guidance: Progress toward Scientific Procedures." *School and Society*, 40:537-42; October 27, 1934.
- Max McConn: "How Much Do College Students Learn?" *North American Review*, 232:446-54; November, 1931.
- Max McConn: "Measurement in Educational Experimentation." *Educational Record*, 15:106-19; January, 1934.
- Max McConn: "The Carnegie Foundation's Study of Secondary and Higher Education in Pennsylvania." *Bulletin of the American Association of Collegiate Registrars*, Vol. 5, No. 2, pp. 43-54, 1929.
- Max McConn: "The Cooperative Test Service." *Journal of Higher Education*, 2:225-32; May, 1931.
- Measurement and Guidance of College Students*, American Council on Education, Committee on Personnel Methods. Baltimore: Williams and Wilkins Co., 1933.
- Minnesota University, Committee on Educational Research: *Studies in College Examinations*. Minneapolis: University of Minnesota, 1934.
- National Society for the Study of Education, Thirty-Second Year-

BIBLIOGRAPHY

- book: *The Teaching of Geography*. Bloomington, Illinois: Public School Publishing Co., 1933.
- Verna L. Newsome: "Making English Grammar Function." *English Journal*, 23:48-57; January, 1934.
- C. W. Odell: *Traditional Examinations and New Type Tests*. New York: Century Co., 1928.
- L. J. O'Rourke: *Rebuilding the English-Usage Curriculum to Insure Greater Mastery of Essentials*. Washington, D.C.: The Psychological Institute, 1934.
- Frederic Palmer: "The College Physics Testing Program and Its Significance for Guidance in Secondary Schools." *Educational Record*, 16:82-96; January, 1935.
- C. L. Persing and H. R. Sattley: "Discovering the Reading Interests of Maladjusted Students." *Bulletin of the American Library Association*, January, 1935.
- "Personnel Methods." *Educational Record*, Supplement No. 8, July, 1928.
- Proceedings of the American Philological Association*. Vol. XXX, 1899.
- E. Prokosch: *Deutsche Sprachlehre*. New York: Henry Holt and Co., 1930.
- Reorganization of Mathematics in Secondary Education*. Mathematical Association of America, Inc., 1923.
- Report of the Classical Investigation*, Part I. Princeton: Princeton University Press, 1924.
- Report of the Third Educational Conference*, New York City, November 1-2, 1934. American Council on Education, 744 Jackson Place, Washington, D.C. (Reprinted in part from the *Educational Record* for January, 1935.)
- Report of the Fourth Educational Conference*, New York City, October 31 and November 1, 1935. *Educational Record*, Supplement No. 9, January, 1936.
- H. N. Rivlin: *Functional Grammar*. New York: Bureau of Publications, Teachers College, Columbia University, 1930.
- G. M. Ruch and G. A. Rice: *Specimen Objective Examinations*. Chicago: Scott, Foresman and Co., 1930.
- Russell C. Ross: An Analysis of the Data Secured from the Iowa Academic Test in World History. Unpublished M. A. thesis, State University of Iowa, August, 1932.

BIBLIOGRAPHY

- P. V. Sangren: "Improvement of Reading through the Use of Tests." *Bulletin of Western State Teachers College*, 27, No. 1, Kalamazoo, Michigan: Western State Teachers College, 1931.
- David Segel: *National and State Cooperative High-School Testing Programs*. U.S. Department of the Interior, Office of Education, Bulletin, 1933, No. 9, Washington, D.C.: Government Printing Office, 1933.
- Howard T. Smith and others: *Report of a Study of the Secondary Curriculum*. Milton, Massachusetts: The Secondary Education Board, 1932.
- Max Smith: *The Relationship between Item Validity and Test Validity*. Contributions to Education, No. 621, New York: Teachers College, Columbia University, 1934.
- John M. and Ruth C. Stalnaker: "A 'Construction Shift' English Test." *English Journal* (College Edition), Vol. 24, No. 8, October, 1935.
- Winston B. Stephens: "Tests and Student Guidance." *Junior-Senior High School Clearing House*, 8: 341-46; February, 1934.
- Ruth Strang: *Personal Development and Guidance in College and Secondary School*. New York: Harper and Brothers, 1934.
- Ruth Strang: *The Role of the Teacher in Personnel Work*. New York: Bureau of Publications, Teachers College, Columbia University, 1935.
- Studies in Modern Language Teaching*. Publications of the American and Canadian Committees on Modern Languages, Vol. XVII. New York: The Macmillan Co., 1930.
- J. B. Tharp: "A Test in French Civilization." *French Review*, 8:283-87; March, 1935.
- L. L. Thurstone: *The Reliability and Validity of Tests*. Ann Arbor, Michigan: Edwards Brothers, 1931.
- E. W. Tieg: *Tests and Measurements for Teachers*. Boston: Houghton Mifflin Co.
- M. E. Townsend: *Student Personnel Services in Teacher Training Institutions*. New York: Bureau of Publications, Teachers College, Columbia University, 1932.
- R. W. Tyler: *Constructing Achievement Tests*. Columbus: Ohio State University, 1934. (Reprints from *Educational Research Bulletin*.)

BIBLIOGRAPHY

- R. W. Tyler and others: *Service Studies in Higher Education*. Bureau of Educational Research Monographs, No. 15, Columbus: Ohio State University, 1932.
- Paul F. Voelker and others: "A Program of Demonstration and Research." *Educational Record*, 16:207-16; April, 1935.
- L. A. Webb and others: *High School Curriculum Reorganization*. Ann Arbor, Michigan: North Central Association, 1933.
- Edgar B. Wesley: "Constructing Tests in the Social Studies." University of Iowa Extension Bulletin, No. 310, *Aids for History Teachers*.
- Michael West: *Bilingualism*. Calcutta: Government of India, Central Publication Branch, 1926.
- Lawrence A. Wilkins and others: *Syllabus of Minima in Modern Foreign Languages*. New York: Board of Education, 1931.
- E. G. Williamson: "Estimation versus Measurement of Improvement in English." *School and Society*, 42:159-62; August 3, 1935.
- E. G. Williamson: "Significance for Educational Guidance of Personal Histories." *The School Review*, 44:41-49; January, 1936.
- E. G. Williamson: "The Cooperative Guidance Movement." *School Review*, 43:273-80; April, 1935.
- E. G. Williamson: "University of Minnesota Testing Bureau." *Personnel Journal*, 12:345-55; April, 1934.
- Ben D. Wood: "Basic Considerations in Educational Testing." *Review of Educational Research*, 3:5-20; February, 1933.
- Ben D. Wood: "Coordinated Examining and Testing Programs." *Educational Record*, 15:48-55; January, 1934.
- Ben D. Wood: "Teacher Selection." Proceedings of the Eleventh Annual Spring Conference of the Eastern-States Association of Professional Schools for Teachers, *Problems in Teacher Training*, Vol. XI (Prentice-Hall), 1936.
- Ben D. Wood: "The Criteria of Individualized Education." *Occupations, The Vocational Guidance Magazine*, 14:781-86; May, 1936.
- Ben D. Wood: "The Major Strategy of Guidance." *Educational Record*, 15:419-44; October, 1934.
- Ben D. Wood: *The New York Experiments with New-Type Modern Language Tests*. Publications of the American and Canadian

BIBLIOGRAPHY

- Committees on Modern Languages, Vol. I. New York: The Macmillan Co., 1927.
- Ben D. Wood: "The Ultimate Basis for Satisfactory College-High School Relations." *Bulletin of the American Association of Collegiate Registrars*, pp. 271-78; July, 1934.
- Ben D. Wood and F. S. Beers: "Knowledge versus Thinking." *Teachers College Record*, 37:487-99; March, 1936.
- Eleanor Perry Wood: "Examining the Uses of Examinations." *Harvard Teachers Record*, 3:59-66; April, 1933.
- Clifford Woody and others: "A Symposium on the Effects of Measurement on Instruction." *Journal of Educational Research*, 28:481-527; March, 1935.
- C. D. Zdanowicz: "Restatement of the Language Requirements for the B.A. Degree in Terms of Attainment." *Modern Language Journal*, 15:354-59; February, 1931.
- George F. Zook: "Accreditation of Secondary Schools in the Light of the North Central Association Report." *Educational Record*, 16:70-81; January, 1935.
- George F. Zook: "Accrediting Schools and Colleges," in *Educational Measurement and Guidance*, American Council on Education, 744 Jackson Place, Washington, D.C., 1933.
- State Testing Programs.* Information about the following state testing programs may be secured from the indicated persons:
- Alabama: W. L. Spencer, Department of Education, Montgomery
- Colorado: A. A. Brown, Fort Morgan, Colorado
- Georgia: F. S. Beers, University of Georgia, Athens
- Indiana: H. H. Remmers, Purdue University, Lafayette
- Iowa: E. F. Lindquist, State University of Iowa, Iowa City
- Kentucky: J. B. Miner, University of Kentucky, Lexington
- Minnesota: E. G. Williamson, University of Minnesota, Minneapolis
- New Mexico: J. C. Knode, University of New Mexico, Albuquerque
- North Carolina: M. R. Trabue, University of North Carolina, Chapel Hill
- Ohio: H. A. Toops, Ohio State University, Columbus
- Texas: H. T. Manuel, University of Texas, Austin

BIBLIOGRAPHY

South Carolina: W. C. McCall, University of South Carolina,
Columbia

Wisconsin: T. L. Torgerson, University of Wisconsin, Madison

Periodicals which frequently include articles on testing and guidance are:

The Educational Record, American Council on Education,
Washington, D.C.

Occupations, The Vocational Guidance Magazine, National
Occupational Conference, 551 Fifth Avenue, New York City

The Journal of Higher Education, Ohio State University,
Columbus

Teachers College Record, Bureau of Publications, Teachers
College, Columbia University, New York City

Journal of Educational Psychology, Warwick and York, Balti-
more.

Review of Educational Research, American Educational Re-
search Association, 1201 Sixteenth Street Northwest,
Washington, D.C.

LATIN ACHIEVEMENT TESTS *

- Bacon, F. Niles. *Diagnostic Tests in Latin*. Based on Gray and Jenkins, "Latin for Today," first year course. New York: Ginn and Co.
- Cooperative Latin Tests*: Junior Form 1933, by W. L. Carr and G. R. Humphries. Form 1933, by J. C. Kirtland, Ruth B. McJimsey, and B. M. Allen. 500 W. 116th St., New York: The Cooperative Test Service of the American Council on Education.
- Davis, S. B., and Hicks, E. E. *Historical Content and Background of Caesar's Gallic War*. Pittsburgh, Pa.: University of Pittsburgh.
- DeFerrari, Roy J., and Foran, T. G. *Deferarri-Foran Latin Tests*. 1) Vocabulary, 2) Forms, 3) Comprehension. Washington, D. C.: The Catholic Education Press.
- Godsey, Edith R. *Diagnostic Latin Composition Forms*, 1 and 2. Yonkers-on-Hudson, N.Y.: World Book Co.
- Henmon, V. A. C. *Henmon Latin Tests*. Forms 1, 2, 3, 4. Yonkers-on-Hudson, N.Y.: World Book Co.
- Hutchinson, Mark E. *Hutchinson Latin Grammar Scale*. Scales A and B. Bloomington, Ill.: Public School Publishing Co.
- Indiana Latin Achievement Tests*. Part I, Vocabulary; Part II, Translation; Part III, Derivatives. Bloomington, Ind.; Bureau of Cooperative Research, School of Education, Indiana University.
- Inglis, Alexander. *Harvard Latin Test on Morphology*. Form A. New York: Ginn and Co.
- Inglis, Alexander. *Harvard Latin Test on Syntax*. Form A. New York: Ginn and Co.
- Inglis, Alexander. *Harvard Latin Vocabulary Tests*. Forms A, B, C, D. New York: Ginn and Co.
- Lohr, L. L., and Latshaw, Harry F. *Lohr-Latshaw Form Test*. Chapel Hill, N.C.: Bureau of Educational Research, School of Education, University of North Carolina.
- Messenger, Wilfrida J. *My Progress Book in Latin*. Columbus, O.: Looseleaf Education, Inc.
- Powers, Francis F. *Powers Diagnostic Latin Test*. Forms 1 and 2.

* Prepared by Professor V. A. C. Henmon.

LATIN ACHIEVEMENT TESTS

1. English-Latin Translation; 2. Nouns and Adjectives; 3. Verbs;
 4. Vocabulary; 5. Comprehension and Syntax, Part A; 6. Comprehension and Syntax, Part B; 7. Derivatives. Bloomington, Ill.: Public School Publishing Co.
- Pressey, L. W. *Pressey Test in Latin Syntax*. Bloomington, Ill.: Public School Publishing Co.
- Stevenson, P. R. *Latin Vocabulary Test*. Bloomington, Ill.: Public School Publishing Co.
- Stevenson, P. R., and Coxe, W. W. *Latin Derivative Test*, Forms I, II, and III. Bloomington, Ill.: Public School Publishing Co.
- Thompson, Harold G. *New York Latin Achievement Test*. Test I, for first semester; Test II, for second semester. Forms A and B. Yonkers-on-Hudson, N.Y.: World Book Co.
- Tyler, Caroline, and Pressey, S. L. *Tyler-Pressey Test in Verb Forms*. Forms 1 and 2. Bloomington, Ill.: Public School Publishing Co.
- Ullman, B. L. and Smalley, A. W. *New Progress Tests in Latin*. New York: The Macmillan Co.
- Ullman, B. L., and Kirby, T. J. *Ullman-Kirby Comprehension Test*. Forms 1 and 2. Iowa City, Ia.: Extension Division, State University of Iowa.
- Ullman, B. L., and Clark, T. L. *Test of Classical References and Allusions*. Iowa City, Ia.: Bureau of Educational Research and Service, State University of Iowa.
- White, Dorrance S. *White Latin Test*. Yonkers-on-Hudson, N.Y.: World Book Co.

PROGNOSIS TEST

- Orleans, Jacob S., and Solomon, Michael. *Orleans-Solomon Latin Prognosis Test*. Yonkers-on-Hudson, N.Y.: World Book Co.

INDEX

- Accrediting of schools and colleges, 465-69
- Achievement tests, characteristics of, 23-24
 - function of, 20, 26
 - in the social studies, 164-65
- Administration time, 116
 - and test results, 101-02
- Aims of instruction in mathematics, 337-38
- Alternate-response test items, 152-58
- Appraisal of teachers by examination results, 463-65
- Appreciation of literature, 387-88
 - essay tests of, 389-92
 - objective tests of, 392-98
- Assigning letter grades, 118-25
- Assimilative objectives in English, 385-410
- Attitudes, the testing of, 209-11
 - toward speaking and writing, 435-37
- Aural tests in modern languages, 302-05
- Average deviation technique for letter grades, 121-25
- Average score, 32, 34

- Baker, Florence M., 301n, 333
- Banc, Loren, 178n
- Boucher, C. S., 460
- Breed, F. S., 324n
- Burnham, Paul S., 268n
- Buswell, G. T., 277n, 312n

- Capitalization tests, 424-25
- Catch questions, 54-55
- Characteristics of achievement tests, 23-24
- Cheydeur, F. D., 294n, 311n
- Chronology exercises, 178-82
- Clarke, Frances M., 302
- Classical Investigation, report of the, 264-66, 268, 273, 276, 279, 280, 282
- Clues, in grammatical structure of item, 67-69
 - in verbal associations, 70-71
- Cole, L. E., 269n
- College Entrance Examination Board, 266-70, 276, 279, 280, 285, 299-300, 454, 461, 462
- Completion exercise, 125-36
 - in mathematics, 374-75
- Composite score, 28
- Composition tests in modern languages, 323-26
- Construction exercises in geometry, 356-60
- "Construction shift" test of English usage, 432-33
- Cook, W. W., 49, 101, 416-19
- Cooperative examination building, 260-61
- Cooperative Test Service, 261
- Copeland, Royal S., 449
- Corell, Alice, 290n
- Correction for guessing, 117
- Crawford, A. B., 268n, 445n
- Creative objectives in English, 410-37
- Critical abilities in English, 386-400
- Cultural element, in Latin, 282-83
 - in modern languages, 308-10
- Cumulative Record Card, 474-75
- Current examinations in modern languages, 292-300

- Dale, Edgar, 328n
- Defects in modern language tests, 294-95
- Denny, E. C., 106
- Derivatives in Latin, 279-82
- Descriptive facts in teacher-made tests, 87
- Diagnostic tests, 22-26

INDEX

- Difficulty of test items, 29-34, 113
 Directions on test blank, 114
 Discriminating power of test item, 41-50
 Distribution of difficulty of items, 31

 Educational Records Bureau, 386n
 Eininger, Hiram J., 185n
 English, evaluation of achievement, 384-437
 objectives, 381-84
 English usage, formal elements of, 410-13
 objective tests of, 415-33
 Essay examinations, and new-type tests, 3, 19
 in literature, 389-91
 in mathematics, 344-45
 in natural sciences, 216-22
 in social studies, 204-09
 Evaluation, of testing techniques, 97-102
 of test items, 50-54
 Examinations as a method of instruction, 458-61
 Experiments in science, 245-47
 Expression, tests for power of, 434-35

 Ford, H. E., 264n, 296, 324
 Foreign language objectives, 12, 264-70, 288-92
 Formal elements of English usage, 410-13
 Function of a general achievement test, 20, 26
 Functioning content of test items, 30, 66-81

 Gates, A. I., 312, 315, 331, 333
 Generalization, in mathematics, 345-50
 in science, 240-45, 247-52
 Geometry, 350-60, 368
 Gilman, G. M., 309
 Godsey, Edith R., 275
 Grading essay examinations, 38, 390-91
 See also Specifications for grading in natural science

 Grammar tests, in Latin, 273-76
 in modern languages, 318-23
 Grammatical usage tests in English, 425-28
 Guessing, correction for, 117
 in recognition tests, 62
 Guidance, the major function of examinations, 469-76

 Haage, Catherine M., 271, 272, 274, 277n
 Hagboldt, Peter, 327n
 Hawkes, H. E., 473n
 Henmon, V. A. C., 264n, 292n, 293n, 295, 301n, 311n, 324n, 490
 Homogeneity, of field tested, 27
 in matching exercises, 151, 173, 375

 Identification of structures in science, 228-30
 Immediate objectives of school examinations, 20
 Improvement of teaching, 461-62
 Incentive to study, examinations as an, 457-58
 Index of discrimination, 49
 Individualized objectives, 13-14
 Information tests in science, 215-24
 Instructional values of test materials in literature, 400
 Insufficient learning and test items, 56-61
 Interest in natural phenomena, 238-39
 Interpretative ideas, importance for test construction, 84-85
 Interpretative test items, 88-89
 Interpreting graphs, 196-97
 Irrelevant cues in test items, 67-71
 Item difficulty, 31-34
 Items employing textbook language, 92-96

 Johnston, J. B., 445n, 457n
 Judd, C. H., 277n, 312n

 Kaufman, F. W., 327n
 Keniston, Hayward, 306

INDEX

- Kinds of examinations, 444-45
 King, Mary H., 201n
 Kirby, T. J., 269n, 277n
 Kurz, H. and G., 309
- Laboratory techniques, 252-55
 Language usage, methods of measuring, 410-33
 Latin, examinations in, 270-83
 objectives in, 264-70
 standardized tests in, 284, 490-91
 Leavitt, Sturgis, 289n
 Letter grades for objective test scores, 118-25
 Library, ability to use, 198-99, 408-10
 Limitations, of achievement tests, 28-29
 of new-type techniques in geometry, 350-60
 of new-type techniques in algebra, 360-62
 See also Essay examinations.
 Limper, Louis H., 328n
 Linguistic knowledge, 326-28
 Literal objectives in English, 385-410
 Literary acquaintance, testing of, 400-05
 Literary history, testing of, 405-07
 Locus problems in geometry, 368
 Logical demonstrations in geometry, 350-56
 Lundeborg, Olav K., 302
- Map reading, 200-02
 Mastery test, 36
 Matching exercises, 147-52
 in mathematics, 375-76
 in social studies, 171-74
 Mathematics, new-type tests, 341-45
 objectives, 337-41
 suggestions for test construction, 345-77
 Modern language, current examination practices, 292-300
 objectives, 288-92
 suggested program of examinations, 301-30
- Multiple-choice exercise, 136-47
 in mathematics, 368-72
 in social studies, 175-78
- Natural science, *see* Science
 Newsome, Verna L., 426n
 New-type and traditional tests in mathematics, 341-45
 New York Regents, 268n, 279, 461, 462
 Non-functioning elements in test items, 73-80
- Objectives, defined in terms of behavior, 10
 for individuals, 13-14
 importance in examination program, 5-7
 in chemistry, 5
 in English, 6, 381-84
 in Latin, 264-70
 in mathematics, 337-41
 in modern languages, 12, 288-92
 in natural sciences, 214
 in zoology, 9
 Oral tests in modern languages, 302-05
 O'Rourke, L. J., 427-28
 Outlining, 194-96
- Paragraph completion exercise, 130-36
 Passing grade, 23, 35-38
 Persing, C. L., 408n
 Philosophy of examinations, 443
 Planning experiments, 245-47
 Power of expression, tests of, 434-35
 Preparation of test copy, 114-15
 Problems of test construction, 18
 Program of examinations in modern languages, 329-30
 Prokosch, E., 327n
 Public education and uniform standards, 455-56
 Punctuation tests, 419-24
 Purin, C. M., 306
 Purposes of achievement tests, 20
- Range of difficulty of items, 31
 Ranking of students, 20

INDEX

- Reading abilities, 386-400
 attitudes, 407-08
 of literature, 387-88
 tests in Latin, 276-79
 tests in modern languages, 311-18
 to locate information, 188-91
 Reasoned understanding, 61, 96
 Reasoning and social studies tests, 183
 Recall exercise, 125-30
 Recognition tests, weaknesses of, 79
 Reference books, ability to use, 198-99, 408-10
 Reproductive objectives in English, 410-37
 Revision of examination materials, 12
 Rice, G. A., 279n, 287n, 318n, 335
 Rivlin, H. N., 426n
 Rogers, Agnes L., 302
 Ross, Russell C., 175
 Rote learning and the new-type tests, 81-96
 Ruch, G. M., 279n, 287n, 318n, 335
 Sampling, 29, 108, 166-69
 Sangren, P. V., 386n
 Sattley, H. R., 408n
 Schindler, A. W., 106
 Science, comprehensive plan of measurement, 258-61
 examinations appropriate for different objectives, 215-57
 types of objectives, 214
 Scientific generalizations, 240-45, 247-52
 Scoring-key, 116
 procedures, 117
 Secondary Education Board, 265n, 270, 276, 279, 325
 Seibert, Louise C., 302
 Self-appraisal, of teachers, 464-65
 of educational institutions, 469
 Sentence completion exercise, 130-36
 Sentence structure tests, 428-29
 Short answer tests, 125-30
 in mathematics, 363-67
 Simple recall exercise, 125-30
 Single score feature in achievement tests, 26
 Smalley, A. W., 275, 278n, 281n, 283n
 Smith, Howard T., 290n
 Social studies, essay type questions, 204-09
 general achievement testing, 163-87
 testing of attitudes, 209-11
 testing of work skills, 187-204
 Sources, of information on science problems, 231-35
 of examination materials in science, 255-57
 Specific determiners in true-false statements, 72
 Specifications, for constructing tests, 108
 for grading in natural science, 218, 220, 221, 222, 233, 244
 Spelling tests, 416-19
 Spread of scores on achievement tests, 31, 33
 Stalnaker, John M. and Ruth C., 432n
 Standardized tests, in Latin, 284, 490-91
 in modern languages, 302
 Standards-enforcement and examinations, 447-57
 Standards used in grading, 37
 Starch, Daniel, 342
 Stoudemire, C. A., 289n
 Subjective evaluation of pupil expression, 414-15
 Summarizing, 192-94
 Table of specifications for constructing test, 108
 Teaching as related to testing, 183-84
 Technical terminology in science, 224-28
 Technical weaknesses of test items, 54-55
 Test items, construction of, 109-11
 Textbook language in test items, 92-96
 Sharp, James B., 264n, 302, 333

INDEX

- | | |
|---|--|
| <p>Thought questions, 61
 Time allowance for test, 116
 True-false statement, 152-58
 in mathematics, 372-74
 True measure of achievement, 29</p> <p>Ullman, B. L., 264n, 275, 277n, 278n,
 281n, 283n
 University of Chicago, 460
 University of Minnesota, 457
 University of Wisconsin, 294
 Unsolved problems in science, 235-
 38
 Usage, objective testing of, 415-33
 subjective evaluation of, 414-15
 Uses of examinations, 3, 20-26, 445-47
 in social studies classes, 184-86</p> <p>Validity, of a test, 21
 of a test item, 39-81
 Van Horne, J. and M., 309</p> | <p>Vocabulary tests, in English 429-30
 in Latin, 270-73
 in modern languages, 305-08
 Voelker, Paul F., 13n</p> <p>Webb, L. A., 290n
 Weighted scores, 117
 Werner, O. H., 298
 Wesley, Edgar B., 146
 West, Michael, 312
 Wilkins, Lawrence A., 290n
 Wood, Ben D., 292n, 293n, 295, 302,
 331n, 470n
 Wood, Eleanor Perry, 457
 Woody, Clifford, 298
 Work skills in the social studies, 187-
 204
 Wrong learning and test items, 62-65</p> <p>Zdanowicz, C. D., 294n
 Zook, George F., 465-69</p> |
|---|--|